DOUBLY ROBUST IDENTIFICATION OF TREATMENT EFFECTS FROM MULTIPLE ENVIRONMENTS

Piersilvio De Bartolomeis
ETH Zurich

Julia Kostin
ETH Zurich

Javier Abad
ETH Zurich

Yixin Wang
University of Michigan

Fanny Yang
ETH Zurich

ABSTRACT

Practical and ethical constraints often dictate the use of observational data for causal inference, particularly in medicine and social sciences. Yet, observational datasets are prone to confounding, potentially compromising the validity of conclusions. While adjusting for all available covariates is a common corrective strategy, this approach can introduce bias in the treatment effect estimates, especially when post-treatment variables are present or some variables remain unobserved. Avoiding this bias often requires detailed knowledge of the underlying causal graph, a challenging and often impractical prerequisite. We propose RAMEN, an algorithm that tackles this challenge by leveraging the heterogeneity of multiple data sources without the need to know or learn the underlying causal graph. Notably, RAMEN achieves *doubly robust identification*: it can identify the treatment effect if either the causal parents of the treatment or those of the outcome are observed. Empirical evaluations across synthetic, semi-synthetic, and real-world datasets show that our approach significantly outperforms existing methods[1].

1 INTRODUCTION

Estimating treatment effects is a key objective in fields such as medicine and social sciences, as it helps determine the impact of interventions like novel treatments or policies. To achieve this goal, researchers often rely on randomized trials since randomizing the treatment assignment guarantees unbiased treatment effect estimates under mild assumptions. However, methods relying on randomized data face several issues, such as small sample sizes, sample populations that do not reflect those seen in the real world, and ethical or financial constraints. Hence, there is growing interest in using observational data to estimate treatment effects when the evidence from randomized data is limited.

A fundamental challenge in using observational data is selecting a *valid adjustment set*, i.e. a set of covariates that can be used to identify and estimate the treatment effect (Hernán & Robins, 2010, Chapter 7). Practitioners often adjust for all available covariates (Austin, 2011), but this approach runs the risk of including *bad controls*—covariates that open backdoor paths between the treatment (T) and the outcome (Y), thereby introducing bias into the treatment effect estimate (Rosenbaum, 1984). For instance, consider the causal graphs illustrated in Figure 1. Is $\{X_1, X_2\}$ always a valid adjustment set? In Figure 1a, $\{X_1, X_2\}$ blocks all backdoor paths between T and Y, whereas including X_1 in Figure 1b opens a backdoor path, introducing bias in the treatment effect estimate. As a result, blindly including all covariates in the adjustment set is not always a valid strategy.

Although the example above might seem artificial, bad controls pose a significant challenge, especially in the social sciences, where the causal ordering of the observed covariates is often not clear (King, 2010; Montgomery et al., 2018). For instance, Acharya et al. (2016) found that up to two-thirds of empirical studies in political science that make causal claims inadvertently include bad

[1]See our GitHub repository: https://github.com/jaabmar/RAMEN/

(a) $\{X_1, X_2\}$ is a valid adjustment set

(b) $\{X_1, X_2\}$ is not a valid adjustment set

Figure 1: Two causal graphs illustrating when the set of all covariates is or is not a valid adjustment set: (a) $\{X_1, X_2\}$ blocks all backdoor paths between T and Y, making it a valid adjustment set; (b) X_1 opens a backdoor path between T and Y, introducing bias in the treatment effect estimate if adjusted for. Unobserved variables are dashed and colored in white, bad controls are colored in red, and good controls are colored in green.

controls in their analysis, leading to biased estimates of the treatment effects. In a first attempt to address this challenge in a data-driven way, Shi et al. (2021) leverage access to multiple heterogeneous data sources—a common scenario in practice, e.g. think of observational studies from different countries—to identify and estimate the treatment effect in the presence of bad controls. However, their approach fails when not all the variables in the causal graph are observed, and the distribution of the unobserved variables shifts across environments, e.g. under the causal graph in Figure 1b.

In this work, we propose **R**obust **A**TE identification from **M**ultiple **EN**vironments (RAMEN), an algorithm that identifies and estimates the average treatment effect in the presence of bad controls without the need to know or learn the complete causal graph. Notably, RAMEN leverages the heterogeneity of multiple data sources to achieve *doubly robust identification*: it can identify the average treatment effect if either the causal parents of the treatment or those of the outcome are observed. In particular, identification is possible even when some variables remain unobserved, e.g. under the causal graph in Figure 1b, where previous methods fail. Our key contributions are outlined below.

- We propose the first algorithm, to our knowledge, that leverages multipleheterogeneous data sources to identify and estimate the average treatment effect in the presence of both bad controls and unobserved variables. Our algorithm is based on a novel double robustness property that offers two strategies to identify the treatment effect: either observe the causal parents of the treatment or those of the outcome.
- We demonstrate that our algorithms significantly outperform existing approaches for treatment effect estimation in the presence of bad controls on synthetic, semi-synthetic and real-world datasets. We further evaluate our method on a real-world example, showing that our results align with established epidemiological knowledge.

2 RELATED WORK

Various criteria and methods have been proposed for the purpose of covariate selection, often in the form of necessary and sufficient conditions for a given causal graph, such as the backdoor criterion and its variations (Pearl, 1995; Shpitser et al., 2010; Vander Weele & Shpitser, 2011; Maathuis & Colombo, 2015; Perković et al., 2018). However, since the causal graph is rarely known in real-world applications, the most common approach assumes that all observed covariates are pre-treatment and includes all of them (Austin, 2011, p. 414). Yet, this strategy has several drawbacks: including certain pre-treatment covariates can introduce M-bias (Entner et al., 2013; Gultchin et al., 2020; Cheng et al., 2022b; Shah et al., 2022), and even when bias is not an issue, selecting a smaller subset of covariates leads to more efficient estimates (Hahn, 2004; White & Lu, 2011; De Luna et al., 2011; Rotnitzky & Smucler, 2020; Witte et al., 2020; Henckel et al., 2022; Guo et al., 2023).

The problems described above are orthogonal to our focus in this paper, which is on scenarios where bad controls are present and the underlying causal graph is unknown (see Appendix B for a complete literature review). In this context, previous works have achieved partial identification, albeit with significant computational costs (Hyttinen et al., 2015; Malinsky & Spirtes, 2017). More recently, several works have proposed methods for point identification under different assumptions. Cheng et al. (2022a) rely on a known anchor variable, Shah et al. (2022) assume all observed covariates are pre-treatment, and Shah et al. (2024) allow for post-treatment covariates but assume all children of the treatment variable are observed and known. In contrast, our work achieves point identification in the presence of bad controls by leveraging multiple heterogeneous data sources and assuming that either the parents of the treatment or the outcome are observed.

Finally, our double robustness property significantly differs from the classic results in the existing literature (Robins et al., 1994; Vansteelandt et al., 2008; Chernozhukov et al., 2018). The key distinction is that our property offers robustness to potential unobserved variables, whereas classic double robustness offers robustness to model misspecification. Some previous work has achieved similar robust identification results in the context of panel data (Arkhangelsky & Imbens, 2022) and instrumental variables (Kang et al., 2016; Hartwig et al., 2017; Guo et al., 2018; Kuang et al., 2020; Hartford et al., 2021)—usually by assuming that only a fraction of the available instruments are valid. However, to our knowledge, we are the first to achieve doubly robust identification in the context of post-treatment and unobserved variables when valid instruments are not available.

3 Problem setting

We assume the data is collected under different experimental conditions, represented by environments $e \in \mathcal{E}$, with $|\mathcal{E}| = n_e$. For each environment $e \in \mathcal{E}$, we have access to a dataset $D^e = \{(X_i, T_i, Y_i)\}_{i=1}^n$ which contains n i.i.d. tuples sampled from the marginal induced by the joint distribution $\mathbb{P}^e$ over (X, U, T, Y). Here, $X \in \mathbb{R}^d$ are the observed covariates, $U \in \mathbb{R}^k$ are the unobserved covariates, $T \in \{0, 1\}$ is a binary treatment assignment variable and $Y \in \mathbb{R}$ is the observed outcome. We denote by $\overline{\mathbb{P}} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathbb{P}^e$ the joint distribution of the pooled environments.

3.1 Causal inference preliminaries

For a fixed directed acyclic graph (DAG) $\mathcal{G}$, we denote the complete set of its nodes by $\tilde{Z}$ and only the *observed* nodes by Z. We denote the index set of parents, ancestors, and descendants for any node $\tilde{Z}_i$ by $\mathrm{Pa}(\tilde{Z}_i)$, $\mathrm{An}(\tilde{Z}_i)$, and $\mathrm{De}(\tilde{Z}_i)$, respectively. Additionally, for an index set S, $\tilde{Z}_S$ denotes the subvector of $\tilde{Z}$ corresponding to the indices in S. In what follows, we formally assume that, for all the experimental settings, the joint distribution $\mathbb{P}^e$ is generated according to a structural causal model (Aldrich, 1989) induced by a DAG $\mathcal{G}$.

Assumption 3.1 (Data distribution)**.** *For each environment $e \in \mathcal{E}$, the distribution $\mathbb{P}^e$ is induced by a structural causal model (SCM), defined as a tuple $\mathcal{M}^e = (\mathcal{G}, \{f_i^e\}_{i=1}^p, \mathbb{P}_\epsilon^e)$ on $p = d + k + 2$ variables $(\tilde{Z}_1, \ldots, \tilde{Z}_p)$, where the observed covariates are $X = \tilde{Z}_{[d]}$, the unobserved covariates are $U = \tilde{Z}_{d+[k]}$, the treatment variable is $T = \tilde{Z}_{p-1}$, and the outcome variable is $Y = \tilde{Z}_p$, with $p \notin \mathrm{An}(T)$. The SCM defines the probability distribution $\mathbb{P}^e$ by setting for each $j \in [p]$*

$$\tilde{Z}_j \leftarrow f_j^e(\tilde{Z}_{\mathrm{Pa}(\tilde{Z}_j)}, \epsilon_j), \quad j = 1, \ldots, p,$$

where $f_j^e : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ is a measurable function and $\epsilon \in \mathbb{R}^p$ is an exogenous noise vector following the joint distribution $\mathbb{P}_\epsilon^e$ over p independent variables.

Further, along the lines of the existing methods in the literature (Shi et al., 2021; Wang et al., 2023), we require the absence of observed mediators between T and Y in the structural causal model.

Assumption 3.2 (Absence of Mediators)**.** *We assume that no observed mediators exist between T and Y, i.e. it holds that*

$$\mathrm{De}(T) \cap \mathrm{An}(Y) \cap [d] = \emptyset.$$

We remark that Assumption 3.2 is *falsifiable* using statistical tests to determine whether a covariate is a mediator between T and Y (see, e.g. Baron & Kenny (1986); Preacher & Hayes (2004)). Further, note that when this assumption is violated, the identified causal quantity corresponds to the natural direct effect (Pearl, 2022), which is still a quantity of separate interest in fields like epidemiology (Tchetgen & VanderWeele, 2014) and social sciences (Imai et al., 2011).

3.2 Treatment effect identification

Our goal is to identify the treatment effects for different environments in the presence of unobserved and post-treatment variables. More specifically, we are interested in the average treatment effects (ATEs) for all environments $e \in \mathcal{E}$, defined as

$$\theta^e = \mathbb{E}_{\mathbb{P}^e}\left[Y^{\mathrm{do}(T=1)} - Y^{\mathrm{do}(T=0)}\right].$$

A common approach for identifying the ATE is to find a *valid adjustment set* (Shpitser et al., 2010), that is, a subset $S \subseteq [d]$ of the observed covariates that satisfies both the classic outcome and treatment identification formulae, i.e. for all environments $e \in \mathcal{E}$ and $t \in \{0, 1\}$ it holds that

$$\mathbb{E}_{\mathbb{P}^e}\left[Y^{\mathrm{do}(T=t)}\right] = \mathbb{E}_{\mathbb{P}^e}\left[\mathbb{E}_{\mathbb{P}^e}[Y \mid X_S, T=t]\right] \text{ and } \mathbb{E}_{\mathbb{P}^e}\left[Y^{\mathrm{do}(T=t)}\right] = \mathbb{E}_{\mathbb{P}^e}\left[\frac{Y\mathbb{I}\{T=t\}}{\mathbb{P}^e(T=t \mid X_S)}\right]. \quad (1)$$

Several criteria have been proposed in the literature to find valid adjustment sets, with the backdoor criterion being the most prominent—see Peters et al. (2017, Sec. 6.6) for a detailed discussion. However, these criteria crucially rely on knowledge of the underlying causal graph—a challenging and often impractical prerequisite. Therefore, it is commonly assumed among practitioners that the set $[d]$ of all observed covariates is a valid adjustment set. This is a reasonable assumption only in settings where all the observed covariates are pre-treatment and there are no unobserved covariates.

In contrast, our work focuses on settings where both post-treatment and unobserved covariates are present. To identify the ATE in such settings, we introduce a key assumption. We allow each environment to have a different joint distribution $\mathbb{P}^e$ over (X, U, Y, T). However, we assume that there exists an invariant node, either T or Y, such that its parents are fully observed and the conditional mean of the node given its parents is invariant across environments.

Assumption 3.3 (Invariant node)**.** *We assume that one of the following holds for all $e \in \mathcal{E}$:*

(a) $\mathrm{Pa}(T)$ *are observed indexes and* $\mathbb{E}_{\mathbb{P}^e}\left[T \mid Z_{\mathrm{Pa}(T)}\right] = \mathbb{E}_{\overline{\mathbb{P}}}\left[T \mid Z_{\mathrm{Pa}(T)}\right], \ \mathbb{P}^e - \text{a.s.}$

(b) $\mathrm{Pa}(Y)$ *are observed indexes and* $\mathbb{E}_{\mathbb{P}^e}\left[Y \mid Z_{\mathrm{Pa}(Y)}\right] = \mathbb{E}_{\overline{\mathbb{P}}}\left[Y \mid Z_{\mathrm{Pa}(Y)}\right], \ \mathbb{P}^e - \text{a.s.}$

We denote the node $V \in \{T, Y\}$ for which the above holds as the ***invariant node*** V_{inv}.

It is worth emphasizing that each of the above assumptions can provide identification of the ATE on its own. Here, we combine these two identification assumptions to obtain doubly robust identification: we only require that either (a) or (b) in Assumption 3.3 holds. This is similar in spirit to the double machine learning literature (Robins & Rotnitzky, 1995; Chernozhukov et al., 2018), where only one of two assumptions about model specification needs to hold to obtain consistent ATE estimates. However, the key difference is that our assumption offers robustness against potential unobserved variables in the underlying causal graph, whereas classic double robustness offers robustness against misspecification of the outcome and treatment functions.

Further, the invariance assumptions (a) and (b) are closely related to the conditions in the invariance-based domain generalization literature, such as Peters et al. (2016); Rojas-Carulla et al. (2018); Gu et al. (2024). While these settings are included in Assumption 3.3 (as we discuss in Appendix A.1), our setting does not require full independence of the noise variable[2], unlike Peters et al. (2016), nor is it limited to the additive noise case, as in Gu et al. (2024), which does not hold in the case of binary treatment variables. Finally, we comment on the observability part of Assumption 3.3: assuming $\mathrm{Pa}(V_{\mathrm{inv}})$ are observed is strictly weaker than causal sufficiency, where the full causal

[2] Although we require independence of exogenous noise variables in Assumption 3.1 for the *full* graph, here, we refer to the graph limited to the observed nodes, where the noise variables ϵ_j can be dependent on $Z_{\mathrm{Pa}(Z_j)}$.

graph is assumed to be observed. Specifically, we allow for some unobserved variables, such as the path $U \to Y$ in Figures 1a and 1b, for which Assumption 3.3 still holds. In contrast, Shi et al. (2021) assume that there are no unobserved variables; as a result, if a parent of Y is unobserved and its distribution shifts across environments, their treatment effect estimate would be biased.

4 METHODOLOGY

In this section, we introduce RAMEN, our method to identify the ATE by leveraging the heterogeneity in the observed data. First, we present a doubly robust population-level estimator and discuss under which conditions it equals to the ATE. Then, we show how to compute this estimator tractably by minimizing a novel invariance loss and propose two algorithms to do so: a combinatorial search over subsets and a more scalable differentiable approach for high-dimensional covariate settings.

4.1 POPULATION-LEVEL ESTIMATOR

In what follows, we denote by $\mathcal{I} = [d] \cup \{p-1, p\}$ the index set corresponding to the observed variables $Z = (X_{[d]}, T, Y)$. For any node $V \in \{T, Y\}$ and any observed subset $S \subseteq \mathcal{I} \setminus V$, we define the conditional means over the pooled and individual environments

$$\overline{m}_S(Z;V) := \mathbb{E}_{\overline{\mathbb{P}}}\left[V \mid Z_S\right] \;\text{ and }\; m^e_S(Z;V) := \mathbb{E}_{\mathbb{P}^e}\left[V \mid Z_S\right].$$

We begin by observing that, by Assumption 3.3, there exists an invariant node $V_{\text{inv}} \in \{T, Y\}$ and a subset of covariates S (given by, e.g, $\text{Pa}(V_{\text{inv}})$), for which the following conditional moment constraint holds for all environments $e \in \mathcal{E}$:

$$\exists S \subseteq \mathcal{I} \setminus V_{\text{inv}} : \;\; m^e_S(Z;V_{\text{inv}}) = \overline{m}_S(Z;V_{\text{inv}}),\; \mathbb{P}^e - \text{a.s.} \tag{2}$$

The set S is not necessarily unique: besides the (observed) parents of V_{inv} for instance, the invariance could also hold for certain supersets of $\text{Pa}(V_{\text{inv}})$. By observing that the conditional moment constraint above is equivalent to the following infinite set of unconditional moment constraints

$$\mathbb{E}_{\mathbb{P}^e}\left[(V_{\text{inv}} - \overline{m}_S(Z;V_{\text{inv}}))h(Z_S)\right] = 0, \text{ for all measurable } h,$$

any set S that satisfies the invariance constraint Equation (2) is also contained in

$$\operatorname*{argmin}_{S \subseteq \mathcal{I} \setminus V_{\text{inv}}} \max_{e \in \mathcal{E}} \left(\sup_{h \in L^0\left(\mathbb{R}^{|S|}\right)} \mathbb{E}_{\mathbb{P}^e}\left[(V_{\text{inv}} - \overline{m}_S(Z;V_{\text{inv}}))h(Z_S)\right] \right)^2 := \operatorname*{argmin}_{S \subseteq \mathcal{I} \setminus V_{\text{inv}}} J_S(Z;V_{\text{inv}}),$$

where $L^0(\mathbb{R}^d)$ denotes the space of measurable functions over $\mathbb{R}^d$. However, since the invariant node V_{inv} is not known beforehand, we search for a set of observed nodes that satisfy the invariance with respect to either T or Y, that is, we want to find

$$S_{\clubsuit} \in \operatorname*{argmin}_{S \subseteq \mathcal{I} \setminus V} \min_{V \in \{T,Y\}} J_S(Z;V). \tag{3}$$

For a minimizer $S_{\clubsuit}$, we then define the corresponding population-level RAMEN, estimator for all environments $e \in \mathcal{E}$ as

$$\theta^e(S_{\clubsuit}) := \mathbb{E}_{\mathbb{P}^e}\left[\bar{\mu}_1(X_{S_{\clubsuit}}) - \bar{\mu}_0(X_{S_{\clubsuit}}) + \frac{(Y - \bar{\mu}_1(X_{S_{\clubsuit}}))T}{\bar{\pi}(X_{S_{\clubsuit}})} - \frac{(Y - \bar{\mu}_0(X_{S_{\clubsuit}}))(1-T)}{1 - \bar{\pi}(X_{S_{\clubsuit}})}\right],$$

where we define the pooled conditional outcome and treatment functions as

$$\bar{\mu}_t(X_{S_{\clubsuit}}) := \mathbb{E}_{\overline{\mathbb{P}}}[Y \mid X_{S_{\clubsuit}}, T = t] \;\text{ and }\; \bar{\pi}(X_{S_{\clubsuit}}) := \mathbb{E}_{\overline{\mathbb{P}}}[T \mid X_{S_{\clubsuit}}].$$

In what follows, we show that under a condition on data heterogeneity, detailed in Assumption 4.1, our population-level estimator $\theta^e(S_{\clubsuit})$ is equivalent for all $S_{\clubsuit}$ that satisfy Equation (3), and it is equal to the true treatment effect. In Sections 4.3 and 4.4, we then discuss how we can construct a good finite-sample ATE estimate in a computationally efficient way.

4.2 DOUBLY ROBUST IDENTIFICATION

Without further assumptions, finding the minimizer of Equation (3) is not sufficient for identifying the ATE: for instance, if there is no variability between distributions $\mathbb{P}^e$, our objective could be trivially minimized by any observed subset S. Only when there is sufficient heterogeneity in the observed environments will $\theta^e(S_{\clubsuit})$ is equivalent to the ATE. We formalize this condition below.

Assumption 4.1 (Identification condition)**.** *For all $V \in \{T, Y\}$ and $S \subset \mathcal{I} \setminus V$, it holds that:*

$$\overline{\mathbb{P}}\left(\overline{m}_S(Z;V) \neq \overline{m}_{\mathrm{Pa}(V)}(Z;V)\right) > 0 \implies \exists e \in \mathcal{E} : \mathbb{P}^e\left(m^e_S(Z;V) \neq \overline{m}_S(Z;V)\right) > 0.$$

Assumption 4.1 can be understood as ensuring that the environments present sufficient heterogeneity. This heterogeneity guarantees that conditioning on any set S with invariant outcome or treatment functions across environments (i.e., a set identified by our method) is equivalent to conditioning on the parents of the invariant node. Although our environment variability assumption is relatively strict, it is a common requirement in the invariance literature (cf. Peters et al. (2016); Arjovsky et al. (2019)). For example, in the simultaneous noise intervention setting described in Peters et al. (2016, Section 4.2.3), Assumption 4.1 can be satisfied with as few as two environments, while in the case of single-node interventions, it requires approximately $\mathcal{O}(p)$ environments.

We now present our formal identification result for the ATE.

Theorem 1 (Doubly robust identification)**.** *Let $S_{\clubsuit}$ be any minimizer of the invariance loss, i.e.*

$$S_{\clubsuit} \in \operatorname*{argmin}_{S \subseteq \mathcal{I} \setminus V} \min_{V \in \{T,Y\}} J_S(Z;V). \tag{4}$$

Then, under Assumptions 3.2,3.3, 4.1, if positivity holds, that is

$$\forall e \in \mathcal{E}: \ \ \mathbb{P}^e(T = t \mid X_{S_{\clubsuit}} = x) > 0, \ \forall t \in \{0,1\} \text{ and } \forall x \in \operatorname{supp}\left(\mathbb{P}^e_X\right),$$

we can identify the treatment effect, i.e. $\forall e \in \mathcal{E}: \ \ \theta^e = \theta^e(S_{\clubsuit})$.

The positivity assumption is widely used for identifying treatment effects in observational studies (Hernán & Robins, 2010, Sec. 3.2). Theorem 1 states that any solution to our invariance loss is a valid adjustment set in the sense that it is sufficient to identify the average treatment effect in all the environments. In what follows, we describe how we can solve the optimization problem in Equation (4) in a computationally efficient way.

4.3 KERNELIZED INVARIANCE LOSS

A major problem of the loss function in Equation (3) is that it is computationally infeasible to search over the entire space of measurable functions. However, we can simplify the problem by restricting h to be in a reproducing kernel Hilbert space (RKHS), and as long as the reproducing kernel of the RKHS is universal (e.g. Gaussian kernel), the two formulations are equivalent (Gretton et al., 2012). More formally, for any subset $S \subseteq \mathcal{I} \setminus V$ and environment $e \in \mathcal{E}$, we can write

$$\begin{aligned}\left(\sup_{h \in L^0\left(\mathbb{R}^{|S|}\right)} \mathbb{E}_{\mathbb{P}^e}\left[\underbrace{\left(V - \overline{m}_S(Z;V)\right)}_{:=\delta_S(Z,V)} h(Z_S)\right]\right)^2 &= \left(\sup_{\|h\|_{\mathcal{H}} \leq 1} \mathbb{E}_{\mathbb{P}^e}\left[\delta_S(Z,V) h(Z_S)\right]\right)^2 \\ &= \left\|\mathbb{E}_{\mathbb{P}^e}\left[\delta_S(Z,V) k(\cdot, Z_S)\right]\right\|^2_{\mathcal{H}} \\ &= \mathbb{E}_{\mathbb{P}^e}\left[\delta_S(Z,V) k\left(Z_S, Z'_S\right) \delta_S(Z',V')\right],\end{aligned}$$

where k is a uniformly bounded reproducing kernel corresponding to a universal RKHS $\mathcal{H}$ (Steinwart, 2001, Definition 4), and (V', Z') is an independent copy of (V, Z) following the same distribution. Hence, we can rewrite our loss in a closed-form solution as

$$J_S(Z;V) = \max_{e \in \mathcal{E}} \mathbb{E}_{\mathbb{P}^e}\left[\delta_S(Z,V) k\left(Z_S, Z'_S\right) \delta_S(Z',V')\right]. \tag{5}$$

Relation to existing invariance losses Starting from the work of Peters et al. (2016), there is considerable literature proposing methods to estimate invariant predictors, especially when the optimal predictor is linear. These methods broadly fall into two categories: hypothesis test-based methods (Peters et al., 2016; Heinze-Deml et al., 2018; Pfister et al., 2019) and optimization-based methods (Arjovsky et al., 2019; Ghassami et al., 2017; Rothenhäusler et al., 2019; 2021; Pfister et al., 2021; Yin et al., 2024; Shen et al., 2023; Gu et al., 2024). Our approach falls in the latter category, with a fundamental distinction. While all these works utilize the invariance principle to improve prediction in unseen environments and generalize to new settings, our goal is to identify a treatment effect in known environments. This is reflected in our loss function, as it does not measure the quality of the predictor in any way—e.g. using a least squares loss. Nonetheless, our invariance loss could also be of interest in the domain generalization literature as it retains the benefits of the invariance loss in Gu et al. (2024) while significantly simplifying their optimization procedure.

4.4 A FULLY DIFFERENTIABLE LOSS

In some cases, searching over all possible subsets of covariates is computationally infeasible. To address this, we propose a continuous relaxation of the optimization problem that can be efficiently solved using gradient descent. Specifically, we select the nodes as $Z_w := B(w) \odot Z$, where the j-th component of $B(w) \in \{0,1\}^{d+1}$ is sampled independently from a Bernoulli distribution with success probability $\mathrm{sigmoid}(w_j)$. We then aim to solve the following optimization problem:

$$w^{\mathrm{opt}} \in \underset{w \in \mathbb{R}^{d+1}}{\mathrm{argmin}} \quad \min_{\theta \in \mathbb{R}^{d+1}, V \in \{T,Y\}} \quad \max_{e \in \mathcal{E}} \quad \mathbb{E}_{\mathbb{P}^e, B(w)} \left[(V - f_\theta(Z_w)) k\left(Z_w, Z'_w\right) \left(V' - f_\theta(Z'_w)\right) \right],$$

where f_θ is a neural network parametrized by θ. Since the weights are discrete, direct differentiation is not possible. To overcome this, we use a Gumbel approximation (Jang et al., 2017; Maddison et al., 2017), where the j-th component of $B(w)$ is approximated as:

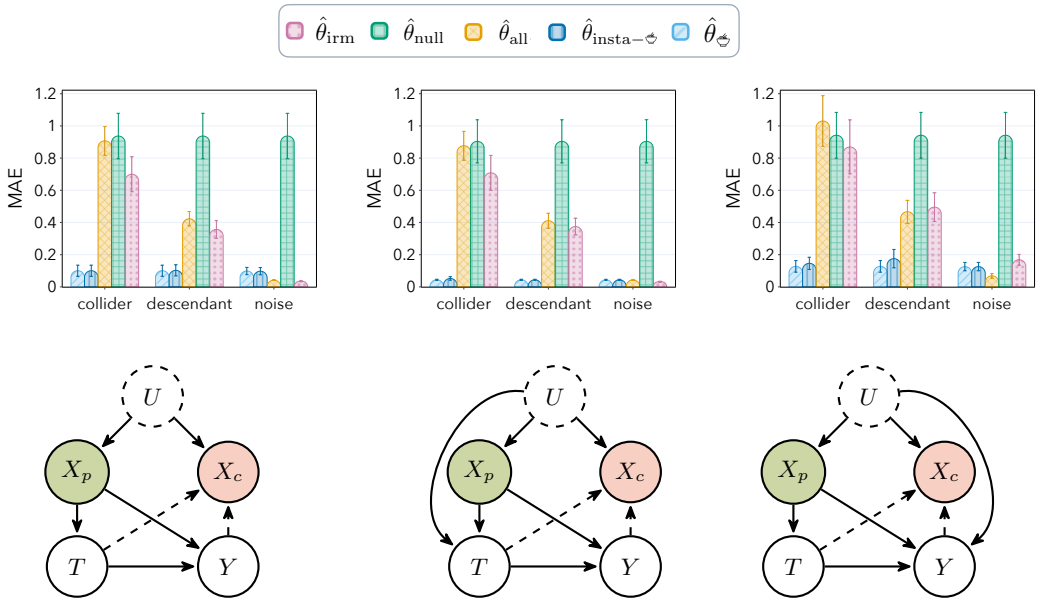
$$B_j(w) \approx \mathrm{sigmoid}\left(\frac{w_j + G_{1,j} - G_{2,j}}{\tau}\right), \quad \text{as} \quad \tau \to 0^+,$$

with $G_{1,j}$ and $G_{2,j}$ being $\mathrm{Gumbel}(0,1)$ random variables. This approximation makes $B(w)$ differentiable (where it was previously discontinuous in w_j), allowing us to optimize using gradient descent while gradually annealing the hyperparameter τ. Finally, we construct the subset of covariates $S_{\text{☕}}$ by including Z_i only if the weights are positive, that is $S_{\text{☕}} = \{i : w_i^{\mathrm{opt}} > 0\}$. We refer the reader to Appendix A.3 for the complete implementation details of our algorithms.

5 EXPERIMENTS

In this section, we evaluate our method through experiments on synthetic, semi-synthetic, and real-world data. We first present experiments on several known DAGs, where the invariances are known and satisfy our assumptions. In line with our theory, RAMEN correctly identifies the ATE, resulting in a low estimation error, whereas other methods tend to fail. We also test RAMEN on a more challenging benchmark by uniformly sampling DAGs using the Erdős–Rényi model—a standard approach for testing causal methods across a wide variety of graph topologies (Huang et al., 2020). Finally, we validate our estimator beyond purely synthetic data: first in a semi-synthetic setting with real-world covariates and then in a real-world setting where we compare the conclusions from RAMEN with established epidemiological findings.

In our experiments, we focus on the statistical task of estimating the average treatment effect (ATE) θ^e for each environment $e \in \mathcal{E}$. To evaluate the performance of an estimator $\hat{\theta}^e$, we compute the mean absolute error (MAE) averaged across environments: $\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} |\theta^e - \hat{\theta}^e|$. We evaluate two implementations of RAMEN: (i) $\hat{\theta}_{\text{☕}}$, based on combinatorial subset search (Section 4.3), and (ii) $\hat{\theta}_{\mathrm{insta-}\text{☕}}$, based on the Gumbel trick (Section 4.4). We compare our algorithms against three baselines: $\hat{\theta}_{\mathrm{irm}}$, the IRM approach for treatment effect estimation proposed by Shi et al. (2021); $\hat{\theta}_{\mathrm{all}}$, which adjusts for all available covariates; and $\hat{\theta}_{\mathrm{null}}$, which does not adjust for any covariates.



(a) T,Y-invariance (b) Y-invariance (c) T-invariance

Figure 2: **(Row 1)** For all the plots: $n = 2500$, $d = 3$, $|\mathcal{E}| = 5$. We plot the mean absolute error averaged across environments when: (a) both invariances are preserved; (b) the invariance w.r.t Y is preserved; (c) the invariance w.r.t T is preserved. We report mean and standard error over 20 runs. **(Row 2)** Graphical models that capture our data generating process: (a) U does not break any invariance; (b) U breaks the invariance between X_p and T; (c) U breaks the invariance between X_p and Y. Dashed nodes are unobserved, and dashed lines denote optional edges.

5.1 Synthetic experiments with known DAGs

We start with data generated from distributions with simple underlying DAGs that satisfy our invariance assumptions, as illustrated in Figure 2 (Row 2). Most importantly, we consider three distinct scenarios[3]: (a) Y and T-invariances, i.e. both (a) and (b) in Assumption 3.3 hold; (b) Y-invariance, i.e. only Assumption 3.3 (b) holds; (c) T-invariance, i.e. only Assumption 3.3 (a) holds. For each of the three different invariance scenarios, we further consider three variants: where X_c is either a descendant of Y, a collider between T and Y, or independent noise. For each of these cases, treatment effects are fixed, and coefficients are sampled from a standard normal distribution. For a description of the complete data-generating process, please refer to Appendix D.1. From theory, we expect that: $\hat{\theta}_{\text{null}}$ should always be biased since there is a confounder between T and Y; $\hat{\theta}_{\text{all}}$ should be biased only when X_c is a collider or a descendant; $\hat{\theta}_{\text{irm}}$ should be biased only in the T-invariance case; $\hat{\theta}_{}$ and $\hat{\theta}_{\text{insta-}}$ should never be biased in these settings.

In Figure 2 (Row 1), we present the empirical MAE for all methods, and we confirm the predictions from theory. Our methods, $\hat{\theta}_{}$ and $\hat{\theta}_{\text{insta-}}$, consistently achieve lower MAE compared to the baselines in all scenarios, indicating that the differentiable relaxation of our method does not significantly compromise statistical performance. Expectedly, for T-invariance, the performance of $\hat{\theta}_{\text{irm}}$ deteriorates markedly—even in scenarios where the post-treatment variable is independent noise, it performs worse than simply adjusting for all available covariates. In contrast, our approach remains robust even when one of the invariances is compromised. Finally, we observe that relying on T-invariance leads to increased error across methods since the adjustment set we recover, the parents of the treatment, is not statistically efficient, see e.g. Henckel et al. (2022, Corollary 3.4).

[3]In Appendix C.2, we also present experiments for cases when none of the invariances hold.

5.2 SYNTHETIC EXPERIMENT WITH RANDOM HIGH DIMENSIONAL DAGS

We randomly draw a graph from the Erdös-Rényi random graph model with a total number of nodes $p = 20$. We do rejection sampling to exclude graphs that either contain mediators—as they violate Assumption 3.2—or do not contain at least a confounder. We then sample data from the resulting DAG via a linear structural causal model, with the only exception being the treatment variable T, which is generated by additionally applying a sigmoid function and then sampling from a Bernoulli distribution. We further post-process the graph, adding a post-treatment variable $X_c = Y + T$ and making unobserved either the parents of T or Y (except common parents), depending on the invariance we want to preserve. We apply a random uniform mean and variance shift to all the nodes in the graph except for T and Y to obtain heterogeneous environments, see Appendix D.1 for further details.

We present here the results for the setting where the parents of T (that are not parents of Y) are unobserved, please refer to Appendix C.3 for additional experiments with the other invariances. We sample 100 different DAGs and vary the number of available environments while keeping the sample size fixed. In Figure 8, we plot the empirical MAE averaged across environments. Notably, we observe that across all settings and numbers of available environments, $\hat{\theta}_{\text{insta-}}$ significantly outperforms all the other baselines. Expectedly, $\hat{\theta}_{\text{irm}}$ fails to surpass all trivial baselines, even with many environments, as it lacks the double robustness property.

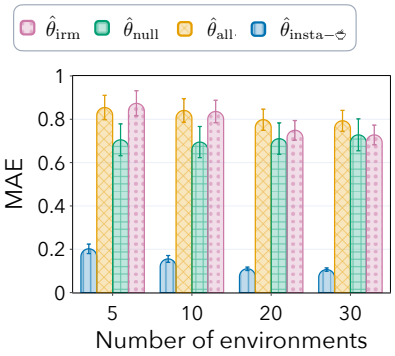


Figure 3: We plot the mean absolute error averaged across environments when the T-invariance is preserved. We sample $n = 2000$ points for each environment; we report mean and standard error over 100 runs.

5.3 SEMI-SYNTHETIC EXPERIMENTS: THE IHDP DATASET

The IHDP dataset contains covariates from $n = 748$ low-birth-weight, premature infants enrolled in a home visitation program designed to improve their cognitive scores (Hill, 2011). Instead of using the commonly adopted synthetic functions from Dorie (2016), we simulate a more challenging non-linear version of the dataset inspired by Kang & Schafer (2007), better reflecting real-world scenarios. Specifically, we retain the 6 continuous features from the original dataset and simulate the outcome Y and treatment assignment T by randomly sampling complex functional forms, such as exponentials and polynomials. In addition, we introduce a 2-dimensional synthetic collider, X_c, as a linear function of T and Y. We generate environments using Gaussian mean shifts in both pre-and post-treatment features, as well as in *either* Y or T, and set the number of environments to $|\mathcal{E}| = 5$. Finally, to make the setting more challenging, we also hide one parent from either Y or T—specifically, from the one that is not invariant. Please refer to Appendix D.2 for additional experimental details.

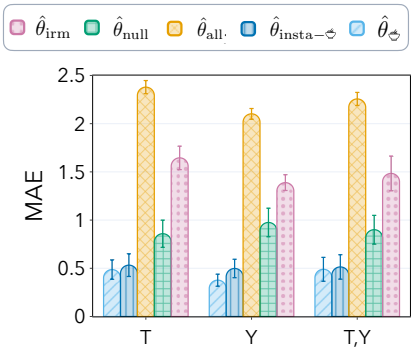


Figure 4: Mean absolute error averaged across environments for the IHDP dataset when different invariances are preserved (T, Y, or both). We consider five environments with $n = 748$ points each; mean and standard error are reported over 20 runs.

Figure 4 presents the results. The increased complexity of the non-linear setup leads to reduced performance across all methods compared to the linear experiments. Despite this, $\hat{\theta}_{}$ and $\hat{\theta}_{\text{insta-}}$ continue to outperform the baselines. Consistent with previous experiments, $\hat{\theta}_{\text{irm}}$ exhibits higher MAE when Y is not invariant across environments, and adjusting for all features ($\hat{\theta}_{\text{all}}$) generally results in poor performance. Interestingly, $\hat{\theta}_{\text{null}}$ performs competitively since the confounders have a limited impact on the outcome and treatment assignment in this dataset. Additional experiments

where the post-treatment feature is either a descendant of the outcome, independent noise, or where neither T nor Y remains invariant are provided in Appendix C.2, along with experiments including mediators between the treatment and the outcome in Appendix C.1.

5.4 Real-world experiment: effect of maternal smoking on birth weight

A classic example of bad controls in epidemiology is the birth-weight paradox: researchers observed higher mortality rates among infants born to smokers compared to non-smokers, but this relationship reversed for low birth-weight infants (Wilcox, 2001). The paradox was resolved by excluding the post-treatment covariate "birth weight" from the adjustment set, which had introduced collider bias (Hernández-Díaz et al., 2006). In our experiments, we analyze data from the same context but instead focus on the effect of smoking on birth weight, as the mortality data is not publicly available.

We evaluate our method on the observational dataset from Cattaneo (2010), which studies the effect of maternal smoking during pregnancy on birth weight ($n = 4642$). We consider 21 covariates from the original dataset, using as treatment T a binary feature indicating whether the mother smoked and, as the outcome Y, the birth weight in grams. The environment is defined by the trimester of birth, and thus $|\mathcal{E}| = 4$. Given the nature of the treatment, we expect that some features are post-treatment, i.e. measured after the mother started smoking, as noted in Wilcox (2001). We provide complete experimental details in Appendix D.3.

Table 1: ATE estimates for the Cattaneo2 dataset using different baselines. We report the mean and standard deviation over 100 initializations of the random seed.

Method	**ATE (mean $\pm$ std)**
$\hat{\theta}_{\text{null}}$	-275.25 ± 10^{-5}
$\hat{\theta}_{\text{all}}$	-157.55 ± 10^{-5}
$\hat{\theta}_{\text{irm}}$	-182.65 ± 48.32
$\hat{\theta}_{\text{insta-☕}}$	-214.60 ± 25.20

Table 1 presents the results of the differentiable version of our method, alongside various baselines. While the ground truth ATE is unknown, the effect estimated by adjusting for the set selected by $\hat{\theta}_{\text{insta-☕}}$ aligns with existing epidemiological literature: both observational and interventional studies (Meyer & Comstock, 1972; Sexton & Hebel, 1984) as well as statistical analyses (Almond et al., 2005; Cattaneo, 2010) estimate a decrease of 200 to 250 grams in birth weight for infants born to smoking mothers compared to non-smoking mothers. In contrast, $\hat{\theta}_{\text{null}}$ overestimates the ATE, whereas both $\hat{\theta}_{\text{all}}$ and $\hat{\theta}_{\text{irm}}$ underestimate it.

6 Discussion and future work

In this work, we proposed **R**obust **A**TE identification from **M**ultiple **EN**vironments (Ramen), a method that leverages multiple environments to identify the ATE in the presence of post-treatment and unobserved variables. To the best of our knowledge, we present the first ATE identification guarantees in this highly relevant, but previously unexplored setting. We introduce a new version of double robustness which concerns identification instead of estimation: we identify the treatment effect if either the causal parents of the treatment or those of the outcome are observed.

Nevertheless, our method presents several limitations. First, like other kernel-based methods, our approach suffers from the curse of dimensionality and the computational complexity associated with kernel matrix computation. Additionally, the requirement for sufficient heterogeneity across environments may be too stringent in some practical cases. Finally, the combinatorial subset is computationally demanding, and the Gumbel trick remains a heuristic solution. Addressing any of these shortcomings would constitute interesting avenues for future work.

Acknowledgements

PDB was supported by the Hasler Foundation grant number 21050. JK was supported by the SNF grant number 204439. JA was supported by the ETH AI Center.

REFERENCES

Avidit Acharya, Matthew Blackwell, and Maya Sen. Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review*, 110(3):512–529, 2016.

John Aldrich. Autonomy. *Oxford Economic Papers*, 41(1):15–34, 1989.

Douglas Almond, Kenneth Chay, and David Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Dmitry Arkhangelsky and Guido Imbens. Doubly robust identification for causal panel data models. *The Econometrics Journal*, 25(3):649–674, 2022.

Peter Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011.

Rina Foygel Barber, Mathias Drton, Nils Sturma, and Luca Weihs. Half-trek criterion for identifiability of latent variable models. *The Annals of Statistics*, 50(6):3174–3196, 2022.

Reuben Baron and David Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173, 1986.

Matias Cattaneo. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155(2):138–154, 2010.

Debo Cheng, Jiuyong Li, Lin Liu, Jixue Liu, Kui Yu, and Thuc Duy Le. Causal query in observational data with hidden variables. *European Conference on Artificial Intelligence*, 2020.

Debo Cheng, Jiuyong Li, Lin Liu, Kui Yu, Thuc Duy Le, and Jixue Liu. Toward unique and unbiased causal effect estimation from data with hidden variables. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):6108–6120, 2022a.

Debo Cheng, Jiuyong Li, Lin Liu, Jiji Zhang, Jixue Liu, and Thuc Duy Le. Local search for efficient causal effect estimation. *IEEE Transactions on Knowledge and Data Engineering*, 2022b.

Debo Cheng, Jiuyong Li, Lin Liu, Jixue Liu, and Thuc Duy Le. Data-driven causal effect estimation based on graphical causal modelling: A survey. *ACM Computing Surveys*, 56(5):1–37, 2024.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Piersilvio De Bartolomeis, Javier Abad, Konstantin Donhauser, and Fanny Yang. Hidden yet quantifiable: A lower bound for confounding strength using randomized trials. *International Conference on Artificial Intelligence and Statistics*, 2024a.

Piersilvio De Bartolomeis, Javier Abad, Konstantin Donhauser, and Fanny Yang. Detecting critical treatment effect bias in small subgroups. *Uncertainty in Artificial Intelligence*, 2024b.

Xavier De Luna, Ingeborg Waernbaum, and Thomas Richardson. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875, 2011.

Ilker Demirel, Edward De Brouwer, Zeshan Hussain, Michael Oberst, Anthony Philippakis, and David Sontag. Benchmarking observational studies with experimental data under right-censoring. *arXiv preprint arXiv:2402.15137*, 2024.

Vincent Dorie. Npci: Non-parametrics for causal inference. 2016. URL https://github.com/vdorie/npci.

Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886, 2011.

Doris Entner, Patrik Hoyer, and Peter Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. *International Conference on Artificial Intelligence and Statistics*, 2013.

Zhuangyan Fang and Yangbo He. Ida with background knowledge. *Uncertainty in Artificial Intelligence*, 2020.

Rina Foygel, Jan Draisma, and Mathias Drton. Half-trek criterion for generic identifiability of linear structural equation models. *The Annals of Statistics*, pp. 1682–1713, 2012.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. *Advances in Neural Information Processing Systems*, 2017.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Yihong Gu, Cong Fang, Peter Bühlmann, and Jianqing Fan. Causality pursuit from heterogeneous environments via neural adversarial invariance learning. *arXiv preprint arXiv:2405.04715*, 2024.

Limor Gultchin, Matt Kusner, Varun Kanade, and Ricardo Silva. Differentiable causal backdoor discovery. *International Conference on Artificial Intelligence and Statistics*, 2020.

Richard Guo, Anton Rask Lundborg, and Qingyuan Zhao. Confounder selection: Objectives and approaches. *arXiv preprint arXiv:2208.13871*, 2022.

Richard Guo, Emilija Perković, and Andrea Rotnitzky. Variable elimination, graph reduction and the efficient g-formula. *Biometrika*, 110(3):739–761, 2023.

Zijian Guo, Hyunseung Kang, Tony Cai, and Dylan Small. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):793–815, 2018.

Jinyong Hahn. Functional restriction and efficiency in causal inference. *The Review of Economics and Statistics*, 86(1):73–76, 2004.

Jason Hartford, Victor Veitch, Dhanya Sridhar, and Kevin Leyton-Brown. Valid causal inference with (some) invalid instruments. *International Conference on Machine Learning*, 2021.

Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, 46(6):1985–1998, 2017.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

Leonard Henckel, Emilija Perković, and Marloes Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):579–599, 2022.

Miguel Hernán and James Robins. Causal inference, 2010.

Sonia Hernández-Díaz, Enrique Schisterman, and Miguel Hernán. The birth weight "paradox" uncovered? *American Journal of Epidemiology*, 164(11):1115–1120, 2006.

Jennifer Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

Zeshan Hussain, Ming-Chieh Shih, Michael Oberst, Ilker Demirel, and David Sontag. Falsification of internal and external validity in observational studies via conditional moment restrictions. *International Conference on Artificial Intelligence and Statistics*, 2023.

Zeshan M Hussain, Michael Oberst, Ming-Chieh Shih, and David Sontag. Falsification before extrapolation in causal effect estimation. *Advances in Neural Information Processing Systems*, 35, 2022.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Do-calculus when the true graph is unknown. *Uncertainty in Artificial Intelligence*, 2015.

Kosuke Imai, Luke Keele, Dustin Tingley, and Teppei Yamamoto. Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(4):765–789, 2011.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *International Conference on Learning Representations*, 2017.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. *Advances in Neural Information Processing Systems*, 31, 2018.

Hyunseung Kang, Anru Zhang, Tony Cai, and Dylan Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

Joseph Kang and Joseph Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, pp. 523–539, 2007.

Rickard KA Karlsson and Jesse H Krijthe. Detecting hidden confounding in observational data using multiple environments. *Advances in Neural Information Processing Systems*, 37, 2023.

Ilmun Kim and Aaditya Ramdas. Dimension-agnostic inference using cross u-statistics. *Bernoulli*, 30(1):683–711, 2024.

Gary King. A hard unsolved problem? post-treatment bias in big social science questions. *Hard Problems in Social Science Symposium*, 2010.

Zhaobin Kuang, Frederic Sala, Nimit Sohoni, Sen Wu, Aldo Córdova-Palomera, Jared Dunnmon, James Priest, and Christopher Ré. Ivy: Instrumental variable synthesis for causal inference. *International Conference on Artificial Intelligence and Statistics*, 2020.

Marloes Maathuis and Diego Colombo. A generalized back-door criterion. *The Annals of Statistics*, 43(3):1060–1088, 2015.

Marloes Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

Chris Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *International Conference on Learning Representations*, 2017.

Daniel Malinsky and Peter Spirtes. Estimating bounds on causal effects in high-dimensional and possibly confounded systems. *International Journal of Approximate Reasoning*, 88:371–384, 2017.

Sarah Mameche, Jilles Vreeken, and David Kaltenpoth. Identifying confounding from causal mechanism shifts. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Mary Meyer and George Comstock. Maternal cigarette smoking and perinatal mortality. *American Journal of Epidemiology*, 96(1):1–10, 1972.

Jacob Montgomery, Brendan Nyhan, and Michelle Torres. How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62 (3):760–775, 2018.

Marco Morucci, Vittorio Orlandi, Harsh Parikh, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. A double machine learning approach to combining experimental and observational data. *arXiv preprint arXiv:2307.01449*, 2023.

Michael Oberst, Alexander D'Amour, Minmin Chen, Yuyan Wang, David Sontag, and Steve Yadlowsky. Understanding the risks and rewards of combining unbiased and possibly biased estimators, with applications to causal inference. *arXiv preprint arXiv:2205.10467*, May 2022.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. Direct and indirect effects. *Probabilistic and causal inference: the works of Judea Pearl*, pp. 373–392, 2022.

Emilija Perković, Markus Kalisch, and Marloes Maathuis. Interpreting and using cpdags with background knowledge. *Uncertainty in Artificial Intelligence*, 2017.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes Maathuis. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18(220):1–62, 2018.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Elements of causal inference: foundations and learning algorithms. *The MIT Press*, 2017.

Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 2019.

Niklas Pfister, Evan Williams, Jonas Peters, Ruedi Aebersold, and Peter Bühlmann. Stabilizing variable selection and regression. *The Annals of Applied Statistics*, 15(3):1220–1246, 2021.

Kristopher Preacher and Andrew Hayes. Spss and sas procedures for estimating indirect effects in simple mediation models. *Behavior research methods, instruments, & computers*, 36:717–731, 2004.

James Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

James Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

James Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89 (427):846–866, 1994.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Paul Rosenbaum. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 147(5): 656–666, 1984.

Paul Rosenbaum and Donald Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *International Conference on Learning Representations*, 2021.

Evan TR Rosenman, Art B Owen, Mike Baiocchi, and Hailey R Banack. Propensity score methods for merging observational and experimental datasets. *Statistics in Medicine*, 41(1):65–86, 2022.

Dominik Rothenhäusler, Peter Bühlmann, and Nicolai Meinshausen. Causal dantzig. *The Annals of Statistics*, 47(3):1688–1722, 2019.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(1): 7642–7727, 2020.

Mary Sexton and Richard Hebel. A clinical trial of change in maternal smoking and its effect on birth weight. *Journal of the American Medical Association*, (7):911–915, 1984.

Abhin Shah, Karthikeyan Shanmugam, and Kartik Ahuja. Finding valid adjustments under non-ignorability with minimal dag knowledge. *International Conference on Artificial Intelligence and Statistics*, 2022.

Abhin Shah, Karthikeyan Shanmugam, and Murat Kocaoglu. Front-door adjustment beyond markov equivalence with limited graph knowledge. *Advances in Neural Information Processing Systems*, 2024.

Uri Shalit, Fredrik Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. *International Conference on Machine Learning*, 2017.

Xinwei Shen, Peter Bühlmann, and Armeen Taeb. Causality-oriented robustness: exploiting general additive interventions. *arXiv preprint arXiv:2307.10299*, 2023.

Claudia Shi, Victor Veitch, and David Blei. Invariant representation learning for treatment effect estimation. *Uncertainty in Artificial Intelligence*, 2021.

Ilya Shpitser, Tyler VanderWeele, and James Robins. On the validity of covariate adjustment for estimating causal effects. *Uncertainty in Artificial Intelligence*, 2010.

Arvid Sjölander. Propensity scores and m-structures. *Statistics in medicine*, 28(9):1416–1420, 2009.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

Ichiro Takeuchi, Quoc Le, Timothy Sears, Alexander Smola, and Chris Williams. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7(7), 2006.

Eric Tchetgen Tchetgen and Tyler VanderWeele. Identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology*, 25(2):282–291, 2014.

Tyler Vander Weele and Ilya Shpitser. A new criterion for confounder selection. *Biometrics*, 67(4): 1406–1413, 2011.

Stijn Vansteelandt, Tyler VanderWeele, Eric Tchetgen Tchetgen, and James Robins. Multiply robust inference for statistical interactions. *Journal of the American Statistical Association*, 103(484): 1693–1704, 2008.

Haotian Wang, Kun Kuang, Haoang Chi, Longqi Yang, Mingyang Geng, Wanrong Huang, and Wenjing Yang. Treatment effect estimation with adjustment feature selection. *Conference on Knowledge Discovery and Data Mining*, 2023.

Luca Weihs, Bill Robinson, Emilie Dufresne, Jennifer Kenkel, Kaie Kubjas Reginald McGee II, McGee II Reginald, Nhan Nguyen, Elina Robeva, and Mathias Drton. Determinantal generalizations of instrumental variables. *Journal of Causal Inference*, 6(1):20170009, 2018.

Halbert White and Xun Lu. Causal diagrams for treatment effect estimation with application to efficient covariate selection. *Review of Economics and Statistics*, 93(4):1453–1459, 2011.

Allen Wilcox. On the importance—and the unimportance—of birthweight. *International journal of epidemiology*, 30(6):1233–1241, 2001.

Janine Witte, Leonard Henckel, Marloes Maathuis, and Vanessa Didelez. On efficient adjustment in causal graphs. *Journal of Machine Learning Research*, 21(246):1–45, 2020.

Lili Wu and Shu Yang. Integrative r-learner of heterogeneous treatment effects combining experimental and observational studies. *Conference on Causal Learning and Reasoning*, 2022.

Shu Yang, Donglin Zeng, and Xiaofei Wang. Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding. *arXiv preprint arXiv:2007.12922*, 2020.

Shu Yang, Chenyin Gao, Donglin Zeng, and Xiaofei Wang. Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):575–596, 04 2023.

Mingzhang Yin, Yixin Wang, and David Blei. Optimization-based causal estimation from heterogeneous environments. *Journal of Machine Learning Research*, 25:1–44, 2024.

APPENDICES

The following appendices provide deferred proofs, experiment details, and ablation studies.

TABLE OF CONTENTS

A Methodology **18**

A.1 Discussion of Assumption 3.3 18

A.2 Proof of Theorem 1 19

A.3 Implementation details 21

A.3.1 Algorithm 1: Combinatorial search over subsets 21

A.3.2 Algorithm 2: Gumbel trick 21

B Extended related work **24**

C Additional experiments **25**

C.1 Robustness to mediators 25

C.2 Robustness to lack of invariance 25

C.3 Additional random graphs experiments 26

C.4 Additional semi-synthetic experiments 26

C.5 Robustness to small sample size 27

C.6 Robustness to violations of Assumption 4.1 27

D Experimental details **29**

D.1 Synthetic experiments 29

D.2 Infant Health and Development Program (IHDP) Dataset 30

D.3 Cattaneo2 32

A METHODOLOGY

A.1 DISCUSSION OF ASSUMPTION 3.3

First, we observe here that Assumption 3.3 is not a minimal "observability" condition on the parents of Y and T: in some cases, it might still be possible to find a valid adjustment set via the observed parents of either T or Y (or both), although no full set of parents was observed (see e.g. Figure 5). However, in such cases, the valid adjustment set or the corresponding regression function cannot be recovered via invariance methods, since neither T nor Y are invariant across environments. Thus, in a way, Assumption 3.3 is a minimal assumption on the DAG if one wants to recover the ATE via invariance of conditional expectations.

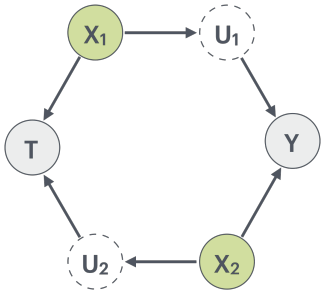


Figure 5: Although neither full set of parents is observed, one can still find a valid adjustment set $\{X_1, X_2\}$ (in green).

Further, we remark that our assumption is neither stronger nor weaker than the commonly used *ignorability* assumption with respect to X (Rosenbaum & Rubin, 1983; Robins & Greenland, 1992). For example, if parents of both T and Y are unobserved, Assumption 3.3 will not hold, but ignorability could still apply if no *common* parent of T and Y is unobserved. Conversely, in graphs with M-bias structures and colliders, the ignorability assumption will not hold.

Below, we give some examples of settings in which Assumption 3.3 holds, including scenarios in the existing invariance literature (Peters et al., 2016; Gu et al., 2024):

Fully observed DAG In Peters et al. (2016), we are given multi-environment data $\{\mathbb{P}^e : e \in \mathcal{E}\}$, where each distribution $\mathbb{P}^e$ is induced by an SCM $\mathcal{M}^e = (\mathcal{G}, \{f_i^e\}_{i=1}^p, \mathbb{P}_\epsilon^e)$, all variables $(X_1, ..., X_d, Y)$ are observed, and for all $\mathbb{P}^e \in \mathcal{E}$ it is satisfied that

$$Y^e = g(X^e_{\mathrm{Pa}(Y)}, \epsilon^e), \epsilon^e \sim F_\epsilon \text{ and } \epsilon^e \perp\!\!\!\perp X^e_{\mathrm{Pa}(Y)}.$$

In particular, the independence condition implies equality of conditional distributions $\mathbb{P}^e(Y|X^e_{\mathrm{Pa}(Y)})$ across environments and thus Assumption 3.3(b).

General DAG with additive noise In Gu et al. (2024), the target variable follows the following data generating process for all $e \in \mathcal{E}$:

$$Y^e = g(X^e_{S^\star}) + \epsilon^e; \quad \mathbb{E}[\epsilon^e | X^e_{S^\star}] = 0,$$

where $S^\star$ is the "true important variable set". This setting is, in a way, more general than Peters et al. (2016), since the noise variable is not required to be independent of the parent variables—instead, the only condition is on the first conditional moment of ϵ^e. Due to the additivity of the noise, Assumption 3.3(b) follows immediately.

General DAG with multiplicative noise We can define a similar setting for multiplicative noise, setting for all $e \in \mathcal{E}$:

$$Y^e = g(X^e_{S^\star})\epsilon^e; \quad \mathbb{E}[\epsilon^e | X^e_{S^\star}] = c,$$

where $S^\star$ is, again, the true parent/important variable set, and c is independent of the environment. We observe that Assumption 3.3(b) follows since it holds that

$$\mathbb{E}^e[Y|X_{S^\star}] = \mathbb{E}^e[g(X_{S^\star})\epsilon|X_{S^\star}] = g(X_{S^\star})\mathbb{E}^e[\epsilon|X_{S^\star}] = cg(X_{S^\star}).$$

General DAG with polynomial noise From the above two examples, it becomes clear that for any $Y = g(X_{S^\star})p_k(\epsilon)$, where p is a polynomial of degree k, we have that Assumption 3.3(b) holds if for all $e \in \mathcal{E}$ it holds that

$$\mathbb{E}^e[\epsilon^{k'} \mid X_{S^\star}] = c_l, \text{ for all } k' \leq k.$$

where $S^\star$ is the important variable/parent set. This condition is strictly weaker than the independence condition since k is finite.

A.2 Proof of Theorem 1

First, we establish that the loss function in Equation (4) attains a value of zero at any minimizer $S_{\clubsuit}$. By Assumption 3.3, there exists an invariant node $V \in \{T, Y\}$ whose parents are observed. Since it holds that $J_{\mathrm{Pa}(V)}(Z; V) = 0$ and $\mathrm{Pa}(V) \subseteq \mathcal{I}$, we conclude that there exists a subset $S \subseteq \mathcal{I}$ such that $\min\{J_S(Z;T), J_S(Z;Y)\} = 0$. Additionally, since the loss function is non-negative, any global minimizer of Equation (4) must have a corresponding loss value of zero.

We now consider two cases, depending on whether the minimum is attained for the node Y or T. We prove our statement for $\mathbb{E}_{\mathbb{P}^e}\left[Y^{\mathrm{do}(T=1)}\right]$; the reasoning is analogous for the control group.

Case 1: $J_{S_{\clubsuit}}(Z;T) = 0$**.** Since we assume that the kernel belongs to a universal RKHS (Steinwart, 2001, Def. 4), it follows from Gretton et al. (2012, Theorem 5), that

$$\forall e \in \mathcal{E}: \;\; \mathbb{E}_{\mathbb{P}^e}[T \mid X_{S_{\clubsuit}}] = \bar{\pi}\left(X_{S_{\clubsuit}}\right), \;\; \mathbb{P}^e - \text{a.s.} \tag{6}$$

Then, by Assumption 4.1, it holds that

$$\bar{\pi}(X_{S_{\clubsuit}}) = \bar{\pi}(Z_{\mathrm{Pa}(T)}), \;\; \overline{\mathbb{P}} - \text{a.s.}$$

We can now identify the treatment effect using the minimizer of the invariance loss as an adjustment set. First, observe that from Equation (6), for any environment $e \in \mathcal{E}$, we have

$$\mathbb{E}_{\mathbb{P}^e}\left[\frac{TY}{\bar{\pi}(X_{S_{\clubsuit}})} + \left(1 - \frac{T}{\bar{\pi}(X_{S_{\clubsuit}})}\right)\bar{\mu}_1(X_{S_{\clubsuit}})\right] = \mathbb{E}_{\mathbb{P}^e}\left[\frac{TY}{\bar{\pi}(Z_{\mathrm{Pa}(T)})}\right] + \mathbb{E}_{\mathbb{P}^e}\left[\left(1 - \frac{T}{\bar{\pi}(Z_{\mathrm{Pa}(T)})}\right)\bar{\mu}_1(X_{S_{\clubsuit}})\right].$$

Now, under the positivity assumption and Assumption 3.3, since the parents of T satisfy the back-door criteria, we can identify the treatment effect, that is, it holds that

$$\mathbb{E}_{\mathbb{P}^e}\left[\frac{TY}{\bar{\pi}(Z_{\mathrm{Pa}(T)})}\right] = \mathbb{E}_{\mathbb{P}^e}\left[Y^{\mathrm{do}(T=1)}\right].$$

It remains to show that the second term is equal to zero:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^e}\left[\left(1 - \frac{T}{\bar{\pi}(X_{S_{\clubsuit}})}\right)\bar{\mu}_1(X_{S_{\clubsuit}})\right] &= \mathbb{E}_{\mathbb{P}^e}\left[\bar{\mu}_1(X_{S_{\clubsuit}})\right] - \mathbb{E}_{\mathbb{P}^e}\left[\frac{T}{\bar{\pi}(X_{S_{\clubsuit}})}\bar{\mu}_1(X_{S_{\clubsuit}})\right] \\
&= \mathbb{E}_{\mathbb{P}^e}\left[\bar{\mu}_1(X_{S_{\clubsuit}})\right] - \mathbb{E}_{\mathbb{P}^e}\left[\mathbb{E}_{\mathbb{P}^e}\left[\frac{T}{\bar{\pi}(X_{S_{\clubsuit}})} \mid X_{S_{\clubsuit}}\right]\bar{\mu}_1(X_{S_{\clubsuit}})\right] \\
&= \mathbb{E}_{\mathbb{P}^e}\left[\bar{\mu}_1(X_{S_{\clubsuit}})\right] - \mathbb{E}_{\mathbb{P}^e}\left[\bar{\mu}_1(X_{S_{\clubsuit}})\right] \\
&= 0,
\end{aligned}$$

where we use the invariance of the conditional expectation $\pi^e(X_{S_{\clubsuit}})$ for $e \in \mathcal{E}$ to show that

$$\mathbb{E}_{\mathbb{P}^e}\left[\frac{T}{\bar{\pi}(X_{S_{\clubsuit}})}\bar{\mu}_1(X_{S_{\clubsuit}}) \mid X_{S_{\clubsuit}}\right] = \bar{\mu}_1(X_{S_{\clubsuit}}).$$

Thus, we conclude that

$$\mathbb{E}_{\mathbb{P}^e}\left[Y^{\mathrm{do}(T=1)}\right] = \mathbb{E}_{\mathbb{P}^e}\left[\frac{TY}{\bar{\pi}(X_{S_{\clubsuit}})} + \left(1 - \frac{T}{\bar{\pi}(X_{S_{\clubsuit}})}\right)\bar{\mu}_1(X_{S_{\clubsuit}})\right].$$

Case 2: $J_{S_{\clubsuit}}(Z;Y) = 0$**.** Again, by the universal property of the kernel k and Gretton et al. (2012, Theorem 5) it follows that

$$\forall e \in \mathcal{E}: \;\; \mathbb{E}_{\mathbb{P}^e}[Y \mid T = t, X_{S_{\clubsuit}}] = \bar{\mu}_t\left(X_{S_{\clubsuit}}\right), \;\; \mathbb{P}^e - \text{a.s.}, \;\; \forall t \in \{0, 1\}.$$

By Assumption 4.1, we have

$$\bar{\mu}_t(X_{S_{\text{☺}}}) = \bar{\mu}_t(Z_{\mathrm{Pa}(Y)}), \;\; \overline{\mathbb{P}} - \text{a.s.}$$

First, observe that

$$\mathbb{E}_{\mathbb{P}^e}\left[\bar{\mu}_1(X_{S_{\text{☺}}})\right] = \mathbb{E}_{\mathbb{P}^e}\left[\bar{\mu}_1(Z_{\mathrm{Pa}(Y)})\right] = \mathbb{E}_{\mathbb{P}^e}\left[Y^{\mathrm{do}(T=1)}\right],$$

since $Z_{\mathrm{Pa}(Y)}$ is a valid adjustment set.

We then show that the second term in our estimand is zero

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^e}\left[\frac{(Y-\bar{\mu}_1(X_{S_{\text{☺}}}))T}{\bar{\pi}(X_{S_{\text{☺}}})}\right] &= \mathbb{E}_{\mathbb{P}^e}\left[\mathbb{E}_{\mathbb{P}^e}\left[\frac{(Y-\bar{\mu}_1(X_{S_{\text{☺}}}))T}{\bar{\pi}(X_{S_{\text{☺}}})} \mid X_{S_{\text{☺}}}\right]\right] \\
&= \mathbb{E}_{\mathbb{P}^e}\left[\frac{1}{\bar{\pi}(X_{S_{\text{☺}}})}\mathbb{E}_{\mathbb{P}^e}\left[(Y-\bar{\mu}_1(Z_{\mathrm{Pa}(Y)}))T \mid X_{S_{\text{☺}}}\right]\right] \\
&= \mathbb{E}_{\mathbb{P}^e}\left[\frac{1}{\bar{\pi}(X_{S_{\text{☺}}})}\left(\mathbb{E}_{\mathbb{P}^e}\left[YT \mid X_{S_{\text{☺}}}\right] - \mathbb{E}_{\mathbb{P}^e}\left[\bar{\mu}_1(Z_{\mathrm{Pa}(Y)})T \mid X_{S_{\text{☺}}}\right]\right)\right] \\
&= \mathbb{E}_{\mathbb{P}^e}\left[\frac{1}{\bar{\pi}(X_{S_{\text{☺}}})}\left(\bar{\mu}_1(X_{S_{\text{☺}}})\mathbb{E}_{\mathbb{P}^e}\left[T \mid X_{S_{\text{☺}}}\right] - \bar{\mu}_1(X_{S_{\text{☺}}})\mathbb{E}_{\mathbb{P}^e}\left[T \mid X_{S_{\text{☺}}}\right]\right)\right] = 0,
\end{aligned}$$

where we have used the invariance of conditional expectations $\mu_1^e(X_{S_{\text{☺}}})$ for $e \in \mathcal{E}$ to show that

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^e}\left[YT \mid X_{S_{\text{☺}}}\right] &= \mathbb{E}_{\mathbb{P}^e}\left[Y \mid X_{S_{\text{☺}}}, T=1\right]\mathbb{P}^e(T=1 \mid X_{S_{\text{☺}}}) \\
&= \mathbb{E}_{\mathbb{P}^e}\left[\mathbb{E}_{\mathbb{P}^e}\left[Y \mid T=1\right] \mid X_{S_{\text{☺}}}\right]\mathbb{P}^e(T=1 \mid X_{S_{\text{☺}}}) \\
&= \mu_1^e(X_{S_{\text{☺}}})\mathbb{P}^e(T=1 \mid X_{S_{\text{☺}}}) = \mu_1^e(X_{S_{\text{☺}}})\mathbb{E}_{\mathbb{P}^e}[T \mid X_{S_{\text{☺}}}] \\
&= \bar{\mu}_1(X_{S_{\text{☺}}})\mathbb{E}_{\mathbb{P}^e}[T \mid X_{S_{\text{☺}}}],
\end{aligned}$$

as well as the fact that for all $e \in \mathcal{E}$, $\bar{\mu}_1(X_{S_{\text{☺}}}) = \bar{\mu}_1(Z_{\mathrm{Pa}(Y)}), \;\; \mathbb{P}^e$-a.s. We now proceed similarly with the remaining two terms: it holds that

$$\mathbb{E}_{\mathbb{P}^e}\left[\bar{\mu}_0(X_{S_{\text{☺}}})\right] = \mathbb{E}_{\mathbb{P}^e}\left[\bar{\mu}_0(Z_{\mathrm{Pa}(Y)})\right] = \mathbb{E}_{\mathbb{P}^e}\left[Y^{\mathrm{do}(T=0)}\right],$$

since $Z_{\mathrm{Pa}(Y)}$ is a valid adjustment set. We now show that the last remaining term is equal to zero. First, we compute

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^e}\left[\frac{(Y-\bar{\mu}_0(X_{S_{\text{☺}}})(1-T)}{1-\bar{\pi}(X_{S_{\text{☺}}})}\right] &= \mathbb{E}_{\mathbb{P}^e}\left[\mathbb{E}_{\mathbb{P}^e}\left[\frac{(Y-\bar{\mu}_0(X_{S_{\text{☺}}})(1-T)}{1-\bar{\pi}(X_{S_{\text{☺}}})} \mid X_{S_{\text{☺}}}\right]\right] \\
= \mathbb{E}_{\mathbb{P}^e}\left[\frac{1}{1-\bar{\pi}(X_{S_{\text{☺}}})}\mathbb{E}_{\mathbb{P}^e}\left[Y-\bar{\mu}_0(X_{S_{\text{☺}}}) - YT + \bar{\mu}_0(X_{S_{\text{☺}}})T \mid X_{S_{\text{☺}}}\right]\right].
\end{aligned}$$

We compute for the inner term that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}^e}\left[Y-\bar{\mu}_0(X_{S_{\text{☺}}}) - YT + \bar{\mu}_0(X_{S_{\text{☺}}})T \mid X_{S_{\text{☺}}}\right] \\
&= \mathbb{E}_{\mathbb{P}^e}\left[Y \mid X_{S_{\text{☺}}}\right] - \bar{\mu}_0(X_{S_{\text{☺}}}) - \mathbb{E}_{\mathbb{P}^e}\left[Y \mid X_{S_{\text{☺}}}, T=1\right]\mathbb{P}^e(T=1 \mid X_{S_{\text{☺}}}) + \bar{\mu}_0(X_{S_{\text{☺}}})\mathbb{P}^e(T=1 \mid X_{S_{\text{☺}}}) \\
&= \mathbb{E}_{\mathbb{P}^e}\left[Y \mid X_{S_{\text{☺}}}, T=1\right]\mathbb{P}^e(T=1 \mid X_{S_{\text{☺}}}) + \mathbb{E}_{\mathbb{P}^e}\left[Y \mid X_{S_{\text{☺}}}, T=0\right]\mathbb{P}^e(T=0 \mid X_{S_{\text{☺}}}) \\
&- \mathbb{E}_{\mathbb{P}^e}\left[Y \mid X_{S_{\text{☺}}}, T=0\right] - \mathbb{E}_{\mathbb{P}^e}\left[Y \mid X_{S_{\text{☺}}}, T=1\right]\mathbb{P}^e(T=1 \mid X_{S_{\text{☺}}}) \\
&+ \mathbb{E}_{\mathbb{P}^e}\left[Y \mid X_{S_{\text{☺}}}, T=0\right]\mathbb{P}^e(T=1 \mid X_{S_{\text{☺}}}) = 0,
\end{aligned}$$

where again we have used the invariance of the conditional expectations $\mu_0^e(X_{S_{\text{☺}}})$ and $\mu_1^e(X_{S_{\text{☺}}})$ across environments.

Combining all four terms, we finally obtain

$$\theta^e = \mathbb{E}_{\mathbb{P}^e}\left[Y^{\mathrm{do}(T=1)}\right] - \mathbb{E}_{\mathbb{P}^e}\left[Y^{\mathrm{do}(T=0)}\right].$$

A.3 IMPLEMENTATION DETAILS

In this section, we describe all the implementation details for our methodology.

Estimation of the loss function We have several choices when it comes to estimating our loss function, as there is a trade-off between statistical and computational efficiency. For instance, one can choose the linear time estimator proposed in Gretton et al. (2012, Section 6) or the efficient estimator proposed in Kim & Ramdas (2024) that runs in quadratic time. In this paper, we estimate

$$\mathbb{H}_e^2(S) := \mathbb{E}_{\mathbb{P}^e}\left[\delta_S(Z;V)k\left(Z_S, \tilde{Z}_S^{\text{os}}\right)\delta_S(\tilde{V}, \tilde{Z}^{\text{os}})\right]$$

using the cross U-statistic from Kim & Ramdas (2024), defined as

$$\hat{\mathbb{H}}_e^2(S) := \frac{2}{n}\sum_{i=1}^{n/2} h_S(Z_i, V_i),$$

$$\text{with } h_S(Z_i, V_i) := \frac{2}{n}\sum_{j=n/2+1}^{n} \delta_S(Z_i, V_i)k(Z_{i,S}, Z_{j,S})\delta_S(Z_j, V_j).$$

Moreover, we would like the two loss functions, i.e., J_Y and J_T, to be on the same scale to avoid any finite sample issues. Therefore, we standardize the cross U-statistic by dividing the empirical variance $\hat{\sigma}\left(\hat{\mathbb{H}}_e^2(S)\right)$, i.e. the finite sample estimate of the variance term

$$\sigma^2\left(\hat{\mathbb{H}}_e^2(S)\right) := \mathbb{E}_{\mathbb{P}^e}\left[\left(h_S(Z) - \mathbb{E}_{\mathbb{P}^e}\left[h_S(Z)\right]\right)^2\right].$$

Choice of kernel An important issue in practice is the selection of the kernel parameters. We used a Gaussian kernel in all of our experiments. We set the bandwidth of the kernel σ to be the median distance between points X in the pooled sample—this remains a heuristic similar to those described in Takeuchi et al. (2006), and the optimum kernel choice is an ongoing area of research.

A.3.1 ALGORITHM 1: COMBINATORIAL SEARCH OVER SUBSETS

We now describe the concrete implementation of our first algorithm.

Since we know that T is a parent of Y, we can simplify our loss function to incorporate this knowledge. Let us define the quantity $\delta_{y,t}(X_S, Y) := Y - \bar{\mu}_t(X_S)$, where $\bar{\mu}_t(X_S) := \mathbb{E}_{\overline{\mathbb{P}}}[Y \mid X_S, T = t]$. We can rewrite the Y-invariance loss function as follows

$$\min_{S\subseteq[d]} \max_{e\in\mathcal{E}, t\in\{0,1\}} \mathbb{E}_{\mathbb{P}^e}\left[\delta_{y,t}(Y, X_S)k\left(X_S, X_S'\right)\delta_{y,t}(Y', X_S') \mid T = t\right].$$

Similarly, we define $\bar{\pi}(X_S) := \mathbb{E}_{\overline{\mathbb{P}}}[T \mid X_S]$ and minimize the T-invariance loss function

$$\min_{S\subseteq[d]} \max_{e\in\mathcal{E}} \mathbb{E}_{\mathbb{P}^e}\left[\delta_t(T, X_S)k\left(X_S, X_S'\right)\delta_t(T', X_S')\right].$$

where $\delta_t(X_S, T) := T - \bar{\pi}(X_S)$.

We explain how to compute the adjustment set explicitly in Algorithm 1, assuming oracle access to the nuisance functions. In practice, nuisance functions can be estimated using the pooled data from all environments.

A.3.2 ALGORITHM 2: GUMBEL TRICK

To deal with the computational infeasibility of searching over all possible subsets of covariates, we propose a continuous relaxation of the optimization problem that can be efficiently solved using gradient descent. The method involves using Gumbel sampling to create differentiable binary masks for covariate selection, which allows optimization via gradient descent. We present the continuous relaxation in Algorithm 2 for obtaining the invariance loss with respect to the node T; the algorithm can be extended analogously to minimize the invariance loss for Y_1 and Y_0.

Algorithm 1 Combinatorial search over subsets ($\hat{\theta}_{\text{☕}}$)

1: **Input:** Data $\{(X_e, Y_e, T_e)\}_{i=1}^{n_e}$, Nuisance functions: $\bar{\pi}$, $\bar{\mu}_0$, $\bar{\mu}_1$
2: **for each subset** $S \subseteq [d]$ **do**
3: **for each environment** $e \in \mathcal{E}$ **do**
4: Compute T-invariance loss using dataset D^e:

$$\hat{J}_S(D^e; T) \leftarrow \frac{4}{n^2} \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} (T_i - \bar{\pi}(X_i)) k(X_{i,S}, X_{j,S}) (T_j - \bar{\pi}(X_j))$$

$$\hat{J}_S(D^e; T) \leftarrow \frac{\hat{J}_S(D^e; Y_0)}{\widehat{\text{Var}}\left(\hat{J}_S(D^e; T)\right)}$$

5: Compute Y-invariance loss using dataset D^e for $T = 1$:

$$\hat{J}_S(D^e; Y_1) \leftarrow \frac{4}{n^2} \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} (Y_i - \bar{\mu}_1(X_i)) k(X_{i,S}, X_{j,S}) (Y_i - \bar{\mu}_1(X_i))$$

$$\hat{J}_S(D^e; Y_1) \leftarrow \frac{\hat{J}_S(D^e; Y_1)}{\widehat{\text{Var}}\left(\hat{J}_S(D^e; Y_1)\right)}$$

6: Compute Y-invariance loss using dataset D^e for $T = 0$:

$$\hat{J}_S(D^e; Y_0) \leftarrow \frac{4}{n^2} \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} (Y_i - \bar{\mu}_0(X_i)) k(X_{i,S}, X_{j,S}) (Y_i - \bar{\mu}_0(X_i))$$

$$\hat{J}_S(D^e; Y_0) \leftarrow \frac{\hat{J}_S(D^e; Y_0)}{\widehat{\text{Var}}\left(\hat{J}_S(D^e; Y_0)\right)}$$

7: **end for**
8: Compute the worst environment losses:

$$\hat{J}_S(T) \leftarrow \max_{e \in \mathcal{E}} J_S(D^e; T), \quad \hat{J}_S(Y_1) \leftarrow \max_{e \in \mathcal{E}} J_S(D^e; Y_1), \quad \hat{J}_S(Y_0) \leftarrow \max_{e \in \mathcal{E}} J_S(D^e; Y_0)$$

9: **end for**
10: **Return:** $S_{\text{☕}} \leftarrow \text{argmin}_S \ \min(\hat{J}_S(T), \hat{J}_S(Y_1), \hat{J}_S(Y_0))$

Algorithm 2 Gumbel trick for subset selection ($\hat{\theta}_{\text{insta}-\text{🍲}}$)

Input: Data $\{(X_e, Y_e, T_e)\}_{e\in\mathcal{E}}$, initial temperature τ_{init}, final temperature τ_{final}, anneal interval k, annealing rate α, learning rates $\eta_{\text{gate}}, \eta_{\text{nn}}$, number of epochs n_{epochs}
Initialize the weights $w_{\bar{\pi}}, w_{\bar{\mu}_0}, w_{\bar{\mu}_1}$ and neural networks $\theta_{\bar{\pi}}, \theta_{\bar{\mu}_0}, \theta_{\bar{\mu}_1}$
Set temperature $\tau \leftarrow \tau_{\text{init}}$
for epoch $= 1$ to n_{epochs} **do**
  if epoch $\bmod k = 0$ **then**
    Update temperature: $\tau \leftarrow \max(\tau_{\text{final}}, \tau \cdot \alpha)$
  end if
  for each environment $e \in \mathcal{E}$ **do**
    for each component j **do**
      Generate mask: $G_{1,j}, G_{2,j} \sim \text{Gumbel}(0, 1)$
      $B_j(w_{\bar{\pi}}) = \text{sigmoid}\left(\frac{w_{\bar{\pi},j}+G_{1,j}-G_{2,j}}{\tau}\right)$
    end for
    Compute invariance loss for T:

$$\hat{J}_{w_{\bar{\pi}}}(D^e;T) \leftarrow \frac{4}{n^2}\sum_{i=1}^{n/2}\sum_{j=n/2+1}^{n}(T_i - f_{\theta_{\bar{\pi}}}(Xi, w_{\bar{\pi}}))k\left(X_{i,w_{\bar{\pi}}}, X_{j,w_{\bar{\pi}}}\right)\left(T_j - f_{\theta_{\bar{\pi}}}(X_{j,w_{\bar{\pi}}})\right)$$

$$\hat{J}_{w_{\bar{\pi}}}(D^e;T) \leftarrow \frac{\hat{J}_{w_{\bar{\pi}}}(D^e;T)}{\widehat{\text{Var}}\left(\hat{J}_{w_{\bar{\pi}}}(D^e;T)\right)}$$

  end for
  Aggregate environment losses:

$$\hat{J}_{w_{\bar{\pi}}}(T) \leftarrow \frac{1}{n_e}\sum_{e\in\mathcal{E}}\hat{J}_{w_{\bar{\pi}}}(D^e;T)$$

  Update gate parameters using gradient descent:

$$w_{\bar{\pi}} \leftarrow w_{\bar{\pi}} - \eta_{\text{gate}}\nabla_{w_{\bar{\pi}}}\hat{J}_{w_{\bar{\pi}}}(T)$$

  Update neural network parameters using gradient descent:

$$\theta_{\bar{\pi}} \leftarrow \theta_{\bar{\pi}} - \eta_{\text{nn}}\nabla_{\theta_{\bar{\pi}}}\hat{J}_{w_{\bar{\pi}}}(T)$$

end for
Return: $S_{\text{🍲}} \leftarrow \{i : w_{\bar{\pi},i} > 0\}$

B EXTENDED RELATED WORK

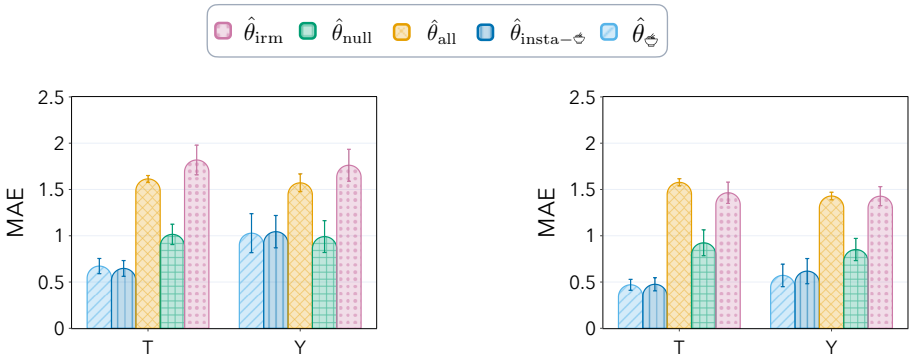
We discuss here the different challenges associated with the problem of selecting covariates for treatment effect identification. Our focus is to highlight differences and similarities with our methodology—we leave out the orthogonal problem of statistical efficiency for the sake of clarity, and we refer the reader to Guo et al. (2022); Cheng et al. (2024) for a complete survey of methods.

Covariate selection with pre-treatment covariates Several works have relaxed the causal sufficiency assumption, allowing for unobserved variables—as long as they are not confounders—while constraining all observed covariates to be pre-treatment. In this setting, the main challenge is M-bias (Sjölander, 2009), which makes adjusting for the full set of covariates not a viable solution. For instance, EHS (Entner et al., 2013) was one of the first methods to obtain partial identification of treatment effects in this setting, however, at the cost of computational inefficiency. Gultchin et al. (2020) propose a more efficient relaxation for EHS to circumvent the computational inefficiency. Further, several more recent works leverage anchor variables to obtain point identification in a computationally efficient way (Cheng et al., 2020; 2022b; Shah et al., 2022). In contrast, our setting is different since we do not assume that all observed covariates are pre-treatment.

Covariate selection under causal sufficiency When all the variables in the causal graph are observed, the only challenge towards identifiability is the presence of post-treatment covariates that can introduce collider bias. Several methods have been proposed to tackle this setting—e.g. IDA (Maathuis et al., 2009) and its variants (Perković et al., 2017; Fang & He, 2020) aim to learn a complete graph from data and then infer a valid adjustment set from it to achieve identifiability. However, they suffer from computational inefficiency since they must first learn the entire causal graph, and they only achieve partial identification. More recently, Shi et al. (2021) consider the setting where multiple environments are available and apply invariant risk minimization (IRM) (Arjovsky et al., 2019) for treatment effect estimation. However, it is widely known that IRM requires many environments—linear in the number of covariates—to generalize well even in the linear regime (Rosenfeld et al., 2021). Finally, Wang et al. (2023) recently proposed a reinforcement learning approach to identify the treatmente effect. In contrast, our approach achieves point identification while being computationally efficient and not requiring causal sufficiency.

Identifiability in linear Gaussian SCMs When the causal graph is a linear Gaussian SCM, the structure of the causal graph (including the hidden variables) imposes algebraic relations between the entries in the covariance matrix and these relations allow or prevent certain aspects of the observable causal model to be recovered. In this regard, several graphical criteria have been identified for deciding whether, in a given causal graph, a specific causal effect can be identified from the covariance for almost all linear Gaussian SCMs compatible with the graph (Drton et al., 2011; Foygel et al., 2012; Weihs et al., 2018; Barber et al., 2022). In contrast, here we are not assuming linearity or gaussianity, and instead rely on multiple heterogeneous data sources for identification.

Combining data from multiple environments Given the challenges associated with estimating treatment effects using non-randomized data, several works propose to detect bias in the treatment effect estimated from observational data by leveraging randomized trials (Yang et al., 2023; Morucci et al., 2023; Hussain et al., 2022; 2023; Demirel et al., 2024; De Bartolomeis et al., 2024b;a), and multiple observational studies (Karlsson & Krijthe, 2023; Mameche et al., 2024). Further, another line of work combines data to estimate the bias and correct for it, ultimately leading to a more accurate treatment effect estimate (Kallus et al., 2018; Rosenman et al., 2022; Wu & Yang, 2022; Yang et al., 2020; Oberst et al., 2022).



(a) With mediator

(b) Without mediator

Figure 6: Mean absolute error averaged across environments for the IHDP dataset with a descendant of the outcome Y when different invariances are preserved (T or Y). We consider the setting with (a) a mediator between T and Y and (b) without a mediator. We consider five environments with $n = 748$ points each; mean and standard error are reported over 20 runs.

C Additional Experiments

C.1 Robustness to Mediators

We study here the robustness of our method to violations of Assumption 3.2. More concretely, we show how the inclusion of a mediator between T and Y affects the ATE estimate for our method and the baselines in several settings. We consider the semi-synthetic experiment setup from Section 5.3, using a 2-dimensional descendant of Y as the post-treatment variable. In Figure 6, we present results for two settings: when the treatment or the outcome is invariant across environments (complete experimental details in Appendix D.2). All baselines show slightly worse performance when a mediator is included. When T is invariant, our method remains competitive and outperforms the baselines, as the parents of T still form a valid adjustment set despite the mediator. However, both $\hat{\theta}_{☕}$ and $\hat{\theta}_{\text{insta-☕}}$ experience a significant drop in performance in the Y-invariance setup, i.e. when T-invariance is violated. This is expected, as in this scenario, we recover the parents of Y, which unfortunately also includes the mediator. A closer inspection of the selected subsets reveals that they usually include the mediator, thus failing to estimate the full effect of T on Y. Instead, our method recovers the natural direct effect of T on Y (Pearl, 2022).

C.2 Robustness to Lack of Invariance

Next, we examine the robustness of our method to violations of the invariance in Assumption 3.3. Specifically, we consider again the semi-synthetic experiments of Section 5.3 in the scenario where neither T- nor Y-invariance holds and there are post-treatment variables (i.e. not the independent noise setting). We provide the results in Figure 7. The performance of our method significantly worsens in this setting, with performance close to the $\hat{\theta}_{\text{null}}$ baseline, as it often recovers the empty set when no invariant node is present. Nonetheless, our method still outperforms $\hat{\theta}_{\text{irm}}$ in all the settings considered.

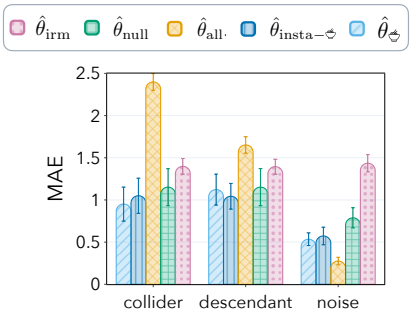
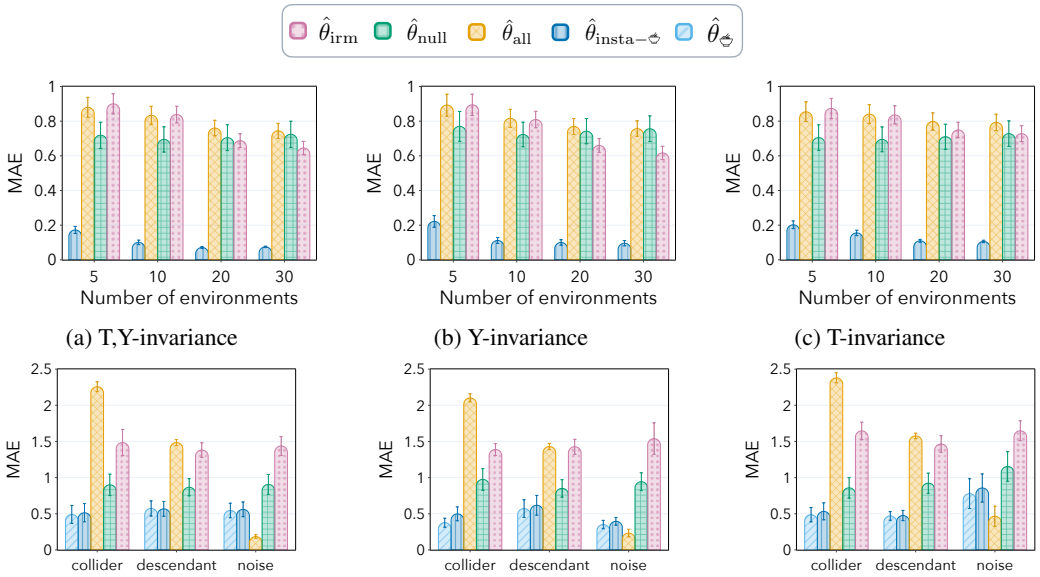


Figure 7: Mean absolute error averaged across environments for the IHDP dataset when no invariance is preserved. We consider five environments with $n = 748$ points each; mean and standard error are reported over 20 runs.



(a) T,Y-invariance (b) Y-invariance (c) T-invariance

(d) T,Y-invariance (e) Y-invariance (f) T-invariance

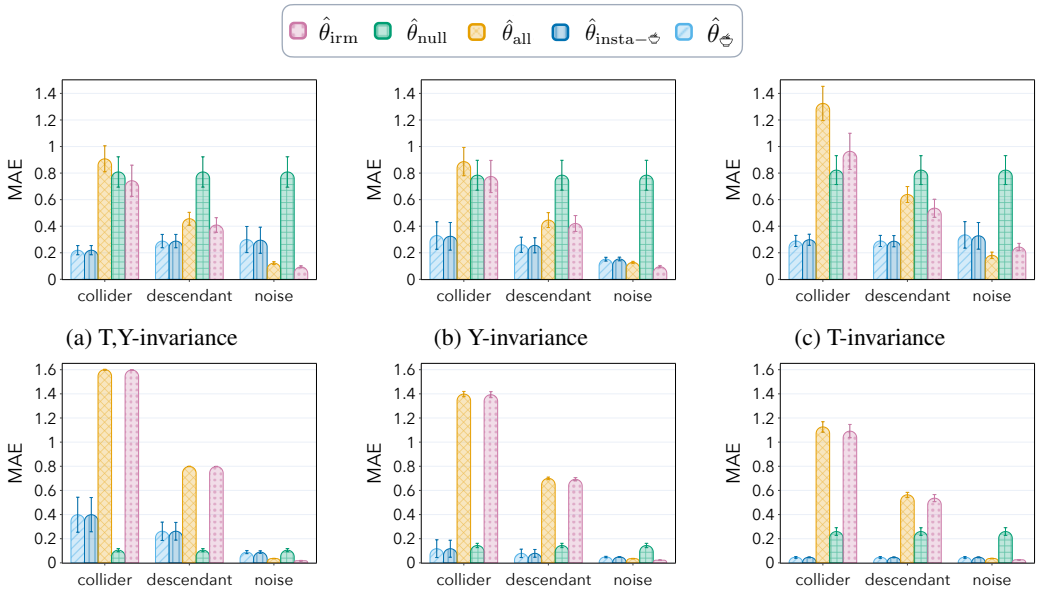
Figure 8: (**Row 1**) We plot the mean absolute error averaged across environments when: (a) no unobserved variables and both invariance w.r.t T and Y are preserved; (b) the parents of T are unobserved but the invariance w.r.t. Y is preserved; (c) the parents of Y are unobserved but the invariance w.r.t T is preserved. For all plots, we sample $n = 2000$ points for each environment; we report mean and standard error over 100 runs. (**Row 2**) Complete experimental results for the semi-synthetic setup described in Section 5.3 using the IHDP dataset. The plots show the mean absolute error averaged across environments for: (a) both T- and Y-invariance, (b) Y-invariance only, (c) T-invariance only.We consider five environments with $n = 748$ points each; mean and standard error are reported over 20 runs.

C.3 Additional random graphs experiments

In Figure 8 (Row 1), we report the MAE averaged across environments for all the invariance settings. First, we observe that across all settings and numbers of available environments, our method significantly outperforms existing baselines. Most notably, $\hat{\theta}_{\text{insta-}}$ achieves relatively small errors even with a limited number of environments. In contrast, $\hat{\theta}_{\text{irm}}$ requires a much larger number of environments to outperform the trivial baselines $\hat{\theta}_{\text{null}}$ and $\hat{\theta}_{\text{all}}$. Further, when the parents of Y are unobserved, $\hat{\theta}_{\text{irm}}$ fails to surpass all trivial baselines, even with many environments—this outcome is expected, as the Y-invariance is broken in this case and $\hat{\theta}_{\text{irm}}$ lacks the double robustness.

C.4 Additional semi-synthetic experiments

We present the complete experimental results using the IHDP dataset (see Section 5.3) in Figure 8 (Row 2). Specifically, we evaluate our proposed method and the baselines under three conditions, where the two-dimensional variable X_c acts as a collider (as described in the main text), descendant, or independent noise. For T-, Y-, and T, Y- invariance, the results align with those obtained in previous sections for linear synthetic experiments. Both $\hat{\theta}_{}$ and its differentiable approximation, $\hat{\theta}_{\text{insta-}}$, outperform the baselines in most settings. The sole exception is when the post-treatment variables are independent noise, where $\hat{\theta}_{\text{all}}$ achieves the best performance. In the case of T-invariance, both our method and $\hat{\theta}_{\text{irm}}$ exhibit slightly worse performance. $\hat{\theta}_{\text{irm}}$ generally underperforms, showing the highest error even under the independent noise setting. The $\hat{\theta}_{\text{null}}$ baseline demonstrates competitive performance overall, likely due to the relatively low influence of confounders in this setup.



(a) T,Y-invariance (b) Y-invariance (c) T-invariance

(d) $\epsilon = 0.0$ (e) $\epsilon = 0.3$ (f) $\epsilon = 0.5$

Figure 9: (**Row 1**) For all the plots: $n = 250$, $d = 3$, $|\mathcal{E}| = 5$. We plot the mean absolute error averaged across environments when: (a) both invariances are preserved; (b) the invariance w.r.t Y is preserved; (c) the invariance w.r.t T is preserved. We report mean and standard error over 20 runs. (**Row 2**) For all the plots: $n = 2500$, $d = 3$, $|\mathcal{E}| = 5$, and only the invariant w.r.t Y is preserved. We plot the mean absolute error averaged across environments for different levels of heterogeneity in the data (higher ϵ corresponds to more heterogeneous data).

C.5 Robustness to small sample size

In Figure 9 (a–c), we report the MAE averaged across environments for the three graphical models introduced in Figure 2 (Row 2). We can observe that our method remains competitive even in the small sample size regime: our method consistently outperforms all baselines when a collider or descendant is present. However, its performance declines in the edge case where the post-treatment variable is independent noise.

C.6 Robustness to violations of Assumption 4.1

We evaluate here the robustness of our method against violations of our identification condition (i.e. Assumption 4.1). To do so, we slightly modify the synthetic experiments presented in Figure 2: We introduce a parameter ϵ^2 to control environment heterogeneity. If $\epsilon^2 = 0$, X has the same distribution across environments. Therefore, there is no heterogeneity across environments, and Assumption 4.1 is violated. On the other hand, if $\epsilon^2 > 0$, the mean and variance of X will shift across environments, with larger shifts as the parameter ϵ^2 increases. Therefore, the heterogeneity across environments increases with ϵ^2, and Assumption 4.1 is more likely to be satisfied.

Example C.1 (Post-treatment variables)**.** *Let $\mathcal{E}$ be the collection of environment indices. For each environment $e \in \mathcal{E}$, we first sample $U \sim \mathcal{N}(0, \epsilon^2 I_{d+2})$. Then, the data is given by*

$$
\begin{aligned}
&X_{p,i} \sim \mathcal{N}(U_i, 0.5 + U_i^2), \;\; \textit{for} \;\; i = 1, \ldots, d; \\
&\quad T \sim \text{Ber}\left(\sigma\left(\beta_t^\top X_p + \epsilon_t\right)\right), \;\; \textit{with} \;\; \beta_t \sim \mathcal{N}(0, I_d) \;\; \textit{and} \;\; \epsilon_t \sim \mathcal{N}(U_{d+1}, 0.5 + U_{d+1}^2); \\
&\quad Y = T + \beta_y^\top X_p + \epsilon_y, \;\; \textit{with} \;\; \beta_y \sim \mathcal{N}(0, I_d) \;\; \textit{and} \;\; \epsilon_y \sim \mathcal{N}(0, 1); \\
&\quad X_c = a \cdot T + b \cdot Y + \epsilon_c, \;\; \textit{with} \;\; \epsilon_c \sim \mathcal{N}(U_{d+2}, 0.5 + U_{d+2}^2).
\end{aligned}
$$

In Figure 9 (d–f), we plot the MAE for different levels of heterogeneity in the data. We can observe that both $\hat{\theta}_{\text{☕}}$ and $\hat{\theta}_{\text{irm}}$ suffer significantly when there is no heterogeneity ($\epsilon = 0.0$). Nevertheless, our method $\hat{\theta}_{\text{☕}}$ consistently outperforms $\hat{\theta}_{\text{irm}}$, even under strong violations of the identification condition. Moreover, $\hat{\theta}_{\text{☕}}$ remains competitive against all baselines when Assumption 4.1 is only weakly satisfied ($\epsilon = 0.3$).

D EXPERIMENTAL DETAILS

Given an adjustment set, we estimate the ATE for each environment $e \in \mathcal{E}$ as follows

$$\hat{\theta}^e_S = \frac{1}{n} \sum_{(X_i,T_i,Y_i)\in D^e} \hat{\mu}^e_S(X_i,1) - \hat{\mu}^e_S(X_i,0) + \frac{(Y_i - \hat{\mu}^e_S(X_i,1))T_i}{\hat{\pi}^e_S(X_i)} - \frac{(Y_i - \hat{\mu}^e_S(X_i,0))(1-T_i)}{1-\hat{\pi}^e_S(X_i)},$$

where $\hat{\mu}^e_S(x,t) = \hat{\mathbb{E}}_{\mathbb{P}^e}[Y \mid T=t, X_S = x]$ and $\pi^e_S(x) = \hat{\mathbb{E}}_{\mathbb{P}^e}[T \mid X_S = x]$.

For $\hat{\theta}_{\text{irm}}$, since the algorithm only learns the outcome function, we estimate the ATE as

$$\hat{\theta}_{\text{irm}} = \frac{1}{n} \sum_{(X_i,T_i,Y_i)\in D^e} \hat{\mu}^e_S(X_i,1) - \hat{\mu}^e_S(X_i,0).$$

D.1 SYNTHETIC EXPERIMENTS

We describe here the data generating process for our synthetic experiments and all the implementation details for the methods.

Example D.1 (Post-treatment variables)**.** *Let $\mathcal{E}$ be the collection of environment indices. For each environment $e \in \mathcal{E}$, we first sample $U \sim \mathcal{N}(0, I_{d+2})$. Then, the data is given by*

$$X_{p,i} \sim \mathcal{N}(U_i, U_i^2), \;\; \textit{for} \;\; i = 1, \dots, d;$$

$$T \sim \text{Ber}\left(\sigma\left(\beta_t^\top X_p + \epsilon_t\right)\right), \;\; \textit{with} \;\; \beta_t \sim \mathcal{N}(0, I_d) \;\; \textit{and} \;\; \epsilon_t \sim \begin{cases} \mathcal{N}(0,1) & \textit{if T is invariant} \\ \mathcal{N}(U_{d+1}, U_{d+1}^2) & \textit{else} \end{cases};$$

$$Y = T + \beta_y^\top X_p + \epsilon_y, \;\; \textit{with} \;\; \beta_y \sim \mathcal{N}(0, I_d) \;\; \textit{and} \;\; \epsilon_y \sim \begin{cases} \mathcal{N}(0,1) & \textit{if Y is invariant} \\ \mathcal{N}(U_{d+1}, U_{d+1}^2) & \textit{else} \end{cases};$$

$$X_c = a \cdot T + b \cdot Y + \epsilon_c, \;\; \textit{with} \;\; \epsilon_c \sim \mathcal{N}(U_{d+2}, U_{d+2}^2).$$

Further, observe that for each choice of invariance, the post-treatment variable X_c can either be a descendant of Y ($a = 0$ and $b = 1$), a collider between T and Y ($a = 1$ and $b = 1$), or independent noise ($a = 0, b = 0$). Finally, under this data-generating process, the average treatment effect is constant across the environments, and it is given by $\theta^e = 1$, for all $e \in \mathcal{E}$.

Random graph data generating process We randomly draw a graph from the Erdös-Rényi random graph model with a density equal to 0.5 and consider graphs with a total number of observed nodes $p = 20$. We do rejection sampling to exclude graphs that either contain mediators (since they violate Assumption 3.2) or do not contain at least a confounder (to make the setting more challenging). We then sample data from the resulting DAG via a linear structural causal model with Gaussian weights using the `causaldag` python library, with the only exception being the treatment variable T, which is generated by additionally applying a sigmoid function and then sampling from a Bernoulli distribution. We further post-process the graph, adding a post-treatment variable $X_c = Y + T$ and removing at random some parents of T or Y depending on which invariance we want to preserve. Therefore, we consider a challenging scenario with both a collider and unobserved variables. To sample data from multiple environments $e \in \mathcal{E}$, within each environment e, we apply a random uniform mean and variance shift to all the nodes in the graph, except for T and Y.

Implementation details We implement our method, $\hat{\theta}_{\text{insta-}☕}$, by performing a hyperparameter search over the following parameters at each iteration: learning rate in the range [0.001, 0.01, 0.1], initial temperature values of [0.5, 0.8, 1.0], and annealing rates of [0.9, 0.95, 0.99]. The optimal combination of these hyperparameters is selected based on minimizing both T-invariance and Y-invariance loss. The outcome functions for $\hat{\theta}_{\text{all}}$, $\hat{\theta}_{☕}$, $\hat{\theta}_{\text{insta-}☕}$ and $\hat{\theta}_{\text{irm}}$ are estimated using a linear regression model. Logistic regression is used for propensity score estimation.

D.2 INFANT HEALTH AND DEVELOPMENT PROGRAM (IHDP) DATASET

The Infant Health and Development Program (IHDP) dataset is a randomized controlled trial focusing on low-birth-weight, premature infants. For our analysis, we keep six continuous covariates from Dorie (2016), representing the child's birth weight, head circumference at birth, number of weeks pre-term, birth order, neonatal health index, and mother's age at birth.

Instead of adopting the treatment and outcome functions from Dorie (2016), we simulate a more challenging scenario inspired by Kang & Schafer (2007). In this setting, each covariate assigned to the treatment (T) or outcome (Y) undergoes a transformation using a predefined set of complex functions similar to those encountered in real-world applications. We introduce the following relationships:

- Confounders: Three of the six covariates are randomly selected to act as confounders, affecting both T and Y.
- Other pre-treatment covariates: The remaining covariates are assigned to affect either T or Y, but not both.
- Post-treatment covariates: We include a two-dimensional post-treatment covariate, denoted as Z, whose generation is detailed below.
- Environmental variation: To introduce variation across environments, we (i) randomly omit a parent of either T or Y and (ii) introduce environment-specific shifts, as detailed below. We apply both to the same node (T or Y) so that the other remains invariant.
- We set ATE $= 2.0$ for all environments.

Modeling of T and Y For each covariate X_i affecting T, we apply a randomly chosen transformation $g_T^{(i)}(x)$ from the following set:

$$g_T^{(i)}(x) \in \left\{ 0.5\log(|x|+1),\ \left(\frac{x}{2}\right)^2,\ x+0.2,\ \exp\left(\frac{x}{2}\right) \right\}.$$

We then compute the logits for the treatment assignment as:

$$T_{\text{logits}} = \sum_i \beta_T^{(i)} g_T^{(i)}(X_i),$$

where $\beta_T^{(i)}$ are coefficients sampled independently from a uniform distribution $\beta_T^{(i)} \sim \mathcal{U}(-0.5, 0.5)$. The binary treatment T is obtained by applying a sigmoid function to T_{logits} and sampling from a Bernoulli distribution:

$$P(T=1) = \sigma(T_{\text{logits}}), \quad T \sim \text{Bernoulli}(P(T=1)),$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function.

Similarly, for each covariate X_j affecting Y, we apply a randomly chosen transformation $g_Y^{(j)}(x)$ from the set:

$$g_Y^{(j)}(x) \in \left\{ 2\log(|x|),\ \left(\frac{x}{2}\right)^2,\ x+1,\ \exp\left(\frac{x}{2}\right) \right\}$$

The outcome Y is then computed as:

$$Y = \sum_j \beta_Y^{(j)} g_Y^{(j)}(X_j),$$

with coefficients $\beta_Y^{(j)}$ sampled from $\beta_Y^{(j)} \sim \mathcal{U}(-2, 2)$.

Incorporating environment-specific shifts To introduce environment-specific variability, we define a hidden variable U that modifies the pre-treatment and post-treatment covariates, outcome, and treatment assignment across different environments. The environments are indexed by $u = 0, 1, 2, 3, 4$. For each environment, we introduce shifts dependent on u.

We first sample coefficients:

$$\beta_{\text{inv}} \sim \mathcal{U}(0.5, 1.0), \quad \beta_X \sim \mathcal{U}(0.5, 1.0).$$

For each environment u, the shifts are generated as:

$$\Delta_{\text{inv}} = u \cdot \beta_{\text{inv}} + \epsilon_{\text{inv}},\ \Delta_X = u \cdot \beta_X + \epsilon_X,\ \Delta_{\text{post}} = u \cdot \beta_X + \epsilon_{\text{post}},$$

where all $\epsilon_{inv}, \epsilon_X, \epsilon_{post}$ are independently sampled from $\mathcal{N}(0, 1)$.

Then, for each environment, the covariates are modified:

$$X = X^0 + \Delta_X,$$

where X^0 represents the original covariate values.

Either Y or T is also shifted, depending on the invariance we aim to preserve:

$$\text{If invariance in } T: \quad Y = Y^0 + \Delta_{\text{inv}}, \quad \text{else if invariance in } Y: \quad T_{\text{logits}} = T^0_{\text{logits}} + \Delta_{\text{inv}},$$

while we add $\mathcal{N}(0, 1)$ to the invariant node.

Generation of post-treatment variables X_c For each environment, we generate a two-dimensional post-treatment variable X_c as follows:

- **Collider:**
$$X_c = Y + T + \epsilon_{\text{post}}, \quad \epsilon_{\text{post}} \sim \mathcal{N}(\Delta_{\text{post}}, I_2).$$
- **Descendant:**
$$X_c = Y + \epsilon_{\text{post}}, \quad \epsilon_{\text{post}} \sim \mathcal{N}(\Delta_{\text{post}}, I_2).$$
- **Independent Noise:**
$$X_c = \epsilon_{\text{post}}, \quad \epsilon_{\text{post}} \sim \mathcal{N}(\Delta_{\text{post}}, I_2).$$

Inclusion of mediators In some settings, we introduce an additional mediator variable influenced by T:

$$\text{Mediator} = \beta_{\text{med}} \cdot T + \epsilon_{\text{med}}, \quad \beta_{\text{med}} \sim \mathcal{U}(-1.0, 1.0), \quad \epsilon_{\text{med}} \sim \mathcal{N}(0, 1).$$

The outcome Y is then adjusted:

$$Y = Y + \text{Mediator}.$$

Summary of data generation process For each environment:

1. Modify covariates: $X = X^0 + \Delta_X$.
2. Compute treatment: $T_{\text{logits}} = \sum_i \beta_T^{(i)}, g_T^{(i)}(X_i) \quad T \sim \text{Bernoulli}(\sigma(T_{\text{logits}}))$.
3. Compute outcome: $Y = \sum_j \beta_Y^{(j)}, g_Y^{(j)}(X_j)$.
4. Apply environmental shift to Y or T and hide a parent of Y or T (we hide the same parent for all environments).
5. Include the ATE $= 2.0$ in the outcome Y
6. If applicable, generate mediator and adjust Y.
7. Generate post-treatment variables Z.

Implementation details We implement our method, $\hat{\theta}_{\text{insta}-\text{☕}}$, by performing a hyperparameter search over the following parameters at each iteration: learning rate in the range [0.001, 0.01, 0.1], initial temperature values of [0.5, 0.8, 1.0], and annealing rates of [0.9, 0.95, 0.99]. The optimal combination of these hyperparameters is selected based on the minimization of both T-invariance and Y-invariance loss. The outcome and treatment assignment functions for both $\hat{\theta}_{\text{all}}$ and $\hat{\theta}_{\text{insta}-\text{☕}}$ are estimated using XGBoost. For these models, we set the number of estimators to 1,000, the learning rate to 0.01, and the maximum tree depth to 6. For the non-linear IRM baseline, we employ the TARNet architecture (Shalit et al., 2017), which consists of a shared representation with a single hidden layer of 200 neurons, followed by two hypothesis-specific hidden layers, each with 100 neurons. Logistic regression is used for propensity score estimation.

D.3 CATTANEO2

The Cattaneo2 dataset (Cattaneo, 2010) studies the effect of maternal smoking on newborn birth weight. We consider 21 covariates, including maternal and paternal age and education, marital status, maternal foreign status, Hispanic origin, alcohol consumption, receipt of prenatal care and the number of prenatal visits, whether the mother had previous children who died, an indicator for low birth weight, months since last birth by the mother, birth month, indicator for whether the baby is first-born, and other variables for which full details are unavailable. The treatment is a binary indicator of smoking status, with 864 mothers in the treatment group and 3,778 in the control group. The outcome is a continuous variable representing birth weight, which we normalize to the interval $[0, 1]$. We exclude the month of birth from the observed features and instead use it to define the environments, creating four environments corresponding to the four quarters of the year.

Implementation details We implement our method, $\hat{\theta}_{\text{insta}-\text{☕}}$, using the following hyperparameters: the number of epochs is set to 700, patience to 100, learning rate to 0.1, initial temperature to 1.0, and annealing rate to 0.9. This configuration was chosen because it provided robust and favorable results across experiments, specifically in minimizing T- and Y-invariance losses. All other hyperparameters are kept from previous experiments. The outcome and treatment assignment functions for both $\hat{\theta}_{\text{all}}$ and $\hat{\theta}_{\text{insta}-\text{☕}}$ are estimated using XGBoost, with the number of estimators set to 1,000, learning rate to 0.01, and maximum depth to 6. For the non-linear IRM implementation, we use the TARNet architecture, as in the IHDP experiments.