Multiply-Robust Causal Change Attribution

Victor Quintas-Martinez [1 2] **Mohammad Taha Bahadori** [3] **Eduardo Santiago** [3] **Jeff Mu** [3] **David Heckerman** [3]

Abstract

Comparing two samples of data, we observe a change in the distribution of an outcome variable. In the presence of multiple explanatory variables, how much of the change can be explained by each possible cause? We develop a new estimation strategy that, given a causal model, combines regression and re-weighting methods to quantify the contribution of each causal mechanism. Our proposed methodology is multiply robust, meaning that it still recovers the target parameter under partial misspecification. We prove that our estimator is consistent and asymptotically normal. Moreover, it can be incorporated into existing frameworks for causal attribution, such as Shapley values, which will inherit the consistency and large-sample distribution properties. Our method demonstrates excellent performance in Monte Carlo simulations, and we show its usefulness in an empirical application. Our method is implemented as part of the Python library `DoWhy` (Sharma & Kiciman, 2020; Blöbaum et al., 2022).

1. Introduction

Analysts are often interested in identifying and quantifying the contribution of multiple possible causes of change to the performance metrics of large-scale systems, such as sales volume, throughput, or retention rate. As an example, a manufacturer compares data from 2023 and 2022 and realizes that net sales have increased. However, many factors have changed between these two time periods, including the characteristics of the product, competitors' prices, or market conditions. How much did each of these variables contribute to the increase in average sales? This is a *change attribution* question. When policymakers and business leaders are interested in using these insights for future changes, the change attribution problem necessarily becomes *causal*.

In the *causal change attribution* problem, we observe two samples of data, including an outcome and multiple explanatory variables. We are also given a causal Directed Acyclic Graph (DAG) that encodes the prior expert knowledge about the causal conditional (in-)dependence relationships among the variables in the data. The objective is to assign a score to each of the causal mechanisms in this DAG that quantifies its contribution to the change in the distribution of the outcome. Although in many instances we observe a change over time, we want to emphasize that our methods apply to more general settings involving a comparison of two samples — for example, differences between groups (e.g., females and males) or geographical locations (e.g., East vs. West Coast of the US).

The contribution of each causal mechanism is generally very difficult to disentangle. We need to characterize what the distribution of the outcome variable would have been if we shifted only some causal mechanisms, but left others unchanged. This is a *counterfactual* distribution, which is not directly observable in the data. We overcome this fundamental challenge by employing a combination of *regression* and *re-weighting* methods. Moreover, our estimating equation is multipy robust, in the sense that it still recovers the parameters of interest even if some components of the model are misspecified. We show that our estimator is consistent and asymptotically normal under weak conditions on the ML algorithms used to learn the regression and the weights. The asymptotic variance is also consistently estimable, allowing us to compute valid standard errors, confidence intervals and p-values. To obtain these results we build on the existing double/debiased ML literature (Chernozhukov et al., 2018a; 2021; 2023).

There is a large body of work on the attribution problem (Efron, 2020; Yamamoto, 2012; Dalessandro et al., 2012), some of which has studied *causal* attribution (Liu et al., 2023; Lu et al., 2023; Mougan et al., 2023; Zhao et al., 2023; Berman, 2018; Ji et al., 2016; Dawid et al., 2014; Shao & Li, 2011). An important branch of the literature has focused on developing suitable definitions of the contribution of each variable, e.g., with variations of Shapley values (Janzing et al., 2024; Chen et al., 2023; Budhathoki et al., 2022; Jung et al., 2022; Sharma et al., 2022) or optimal transport methods (Kulinski & Inouye, 2023), while remaining agnostic about estimation. Our contribution is

[1]MIT Department of Economics. [2]This work was completed while VQM was an intern at Amazon. [3]Amazon. Correspondence to: Victor Quintas-Martinez <vquintas@mit.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

related but distinct: the estimator we develop can be readily integrated into existing frameworks for causal change attribution, such as Shapley values, or it may be of interest in its own right. Shapley values based on our estimator will inherit its consistency and asymptotic normality properties; we also propose a convenient and computationally efficient bootstrap procedure to compute their standard errors.

Although other recent algorithms have been proposed to estimate Shapley Values, a majority of these algorithms are not multiply-robust, and generally they are based on regression only (learning the distribution of the outcome given the explanatory variables).[1] As such, they will not perform well unless we have a high-quality estimator of the regression function. In high-dimensional or non-parametric settings, estimators for those objects will typically exhibit slow convergence rates, rendering traditional normal-based large sample inference (standard errors, confidence intervals or p-values) invalid. We provide the first formal analysis of the asymptotic properties of an estimator in the causal change attribution setting of Budhathoki et al. (2021), which allows us to perform valid inference in large samples (standard errors, confidence intervals and tests).

Causal change attribution is related to the classical problem of treatment effect estimation (Pearl, 2009; Imbens & Rubin, 2015; Peters et al., 2017) in the need to evaluate counterfactual distributions, but there are some fundamental differences. Within the treatment effects literature, the closest problem to ours is that of *causal mediation*. Tchetgen-Tchetgen & Shpitser (2012) gave multiple-robustness results for mediation analysis with a single mediator (explanatory variable). The case with multiple mediators and an arbitrary DAG is substantially more challenging. Although causal mediation with multiple mediators has been studied before (e.g., Daniel et al., 2015), we provide the first multiply-robust estimator in this setting. We also believe to be the first to explicitly build a connection between the causal change attribution problem and the causal mediation literature.

2. Methodology

In this section we describe our multiply-robust methodology for causal change attribution. Section 2.1 defines the causal model and the causal change attribution problem and its connections and contrasts with treatment effect estimation. Section 2.2 gives an identification result in a simple example, which we then generalize in Section 2.3. Section 2.4 describe how to implement our estimator, and 2.5 gives large-sample inference results. Finally, in Section 2.6 we explain how our method can be incorporated within existing frameworks of causal change attribution, such as Shapley values. We provide the technical conditions and proofs for all lemmas and theorems in Appendix A.

2.1. Setting

We observe multiple i.i.d. measurements of the same variables $(T, \boldsymbol{X}, Y)$. Each observation consists of a sample indicator $T \in \{0, 1\}$, K explanatory variables $\boldsymbol{X} := (X_1, \ldots, X_K)$ and an outcome of interest $Y \in \mathcal{Y} \subset \mathbb{R}$. The explanatory variables in $\boldsymbol{X}$ could be continuous, categorical, or even unstructured data types such as text or images.

We assume that the distribution of $(\boldsymbol{X}, Y) \mid T = t$ has a causal Markov factorization (Spirtes et al., 2000):

$$P^{(t)}_{(\boldsymbol{X},Y)} := P^{(t)}_{Y|\boldsymbol{X}} \prod_{k=1}^{K} P^{(t)}_{X_k|\mathrm{PA}_k}, \tag{F}$$

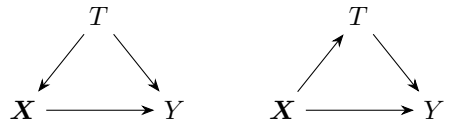
where PA_k are the *parents* (direct causes) of X_k in the underlying causal DAG, other than T. Throughout, the superscript (t) denotes conditioning on $T = t \in \{0, 1\}$. Each conditional distribution on the right-hand side of (F) is called a *causal mechanism* (Peters et al., 2017). Without loss of generality, we assume that the explanatory variables are labeled so that $k < k'$ if $X_k \in \mathrm{PA}_{k'}$, i.e., the causal antecedents of $X_{k'}$ in the DAG have indices lower than k'. In practice, the researcher only needs to know one such *causal ordering*, rather than the full causal graph for $(\boldsymbol{X}, Y)$. Our causal model implies a distribution for $Y \mid T = t$ by marginalization, i.e.,

$$P^{(t)}_{Y} := \int P^{(t)}_{Y|\boldsymbol{X}} \prod_{k=1}^{K} \mathrm{d}P^{(t)}_{X_k|\mathrm{PA}_k}.$$

The problem of *change attribution* tries to quantify how much of the differences between $P^{(1)}_Y$ and $P^{(0)}_Y$ are due to shifting each causal mechanism from its distribution at $T = 0$ to its distribution at $T = 1$. To clarify, intervening on a causal mechanism $P_{X_k|\mathrm{PA}_k}$ may impact the outcome both directly and indirectly through other explanatory variables which are causal descendants of X_k. We are interested on the total effect of changing $P_{X_k|\mathrm{PA}_k}$, without fixing the marginal distribution of its causal descendants. Finally, note that we allow the causal mechanisms to differ between samples, but the underlying DAG is assumed to be the same.

We want to draw a clear distinction between attribution and a related problem that has received a much larger attention in the literature, Average Treatment Effect (ATE) estimation under unconfoundedness. The difference between the

[1] The only exception that we are aware of is Jung et al. (2022). However, their notion of *do*-Shapley values decomposes the effect of fixing a value for the explanatory variables, rather than a distribution for each causal mechanism, as we do in this paper. They consider only discrete-valued covariates, and their approach does not seem easily generalizable beyond that.



(a) Causal Change Attribution. (b) ATE under Unconfoundedness.

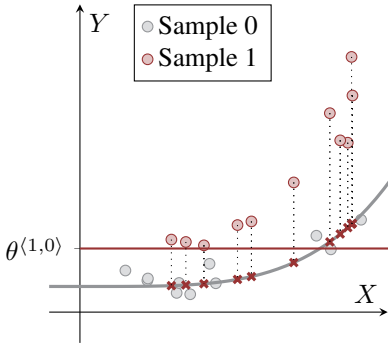
Figure 1. Two possible DAGs for $(T, \boldsymbol{X}, Y)$.

causal models underlying these two problems is depicted in Figure 1. In an ATE setting, $\boldsymbol{X}$ typically represents pre-treatment covariates, that affect both the outcome Y and the propensity to receive a binary treatment T. To compute the causal effect of T we need to control for those covariates, that is, compare units with similar values of $\boldsymbol{X}$. In contrast, causal change attribution acknowledges that both the distributions of $\boldsymbol{X}$ and of Y given $\boldsymbol{X}$ vary depending on whether $T = 0$ or $T = 1$. The goal of causal change attribution is to quantify how much of the effect of T on Y goes through (is mediated by) each causal mechanism in this distribution. Under an unconfoundedness assumption about the assignment of T, the problem of causal attribution can be thought of as decomposing the ATE of T on Y into the Natural Direct Effect and the Natural Indirect Effect corresponding to each causal mechanism (Pearl, 2009, §4.5; Daniel et al., 2015). We describe this in more detail, including the corresponding structural equation model and potential outcomes, in Appendix B.

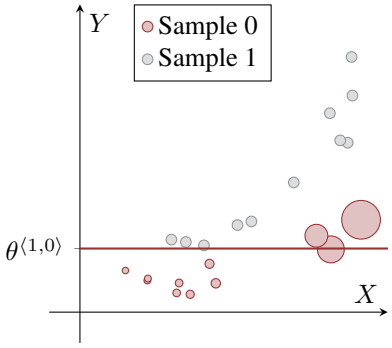
We note, however, that our definition of attribution does not require T to be a treatment that can be administered in the "interventional" sense. As discussed, T could also be an indicator for group membership (e.g., female or male), time period (e.g., before and after a certain date), or geographical location (e.g., East vs. West Coast of the US). Under which conditions can this change attribution be interpreted as *causal*? We will assume that the researcher knows a causal ordering of the underlying true causal graph for $(\boldsymbol{X}, Y)$. This will allow us to quantify the contribution of each *causal mechanism*, rather than changes in the *marginal distributions* of covariates. In practice, researchers can use a combination of causal discovery methods (e.g., the conditional independence test of Zhang et al., 2011) and domain knowledge to obtain the DAG. We also assume that there is no unobserved variable U such that $T \to U$, $U \to \boldsymbol{X}$ and $U \to Y$. We discuss ways to relax this assumption in future research in Section 4.

2.2. Preliminary Example

We begin with a simple yet illustrative example. Consider the simplest situation of $K = 1$. How much of the difference in means $\mathrm{E}[Y \mid T = 1] - \mathrm{E}[Y \mid T = 0]$ is due to changing P_X? How much of it is due to changing $P_{Y|X}$?



(a) Regression: We learn the dependence between Y and X in sample 0 through the regression function $\gamma(X)$ (gray curve). Subsequently, we average the fitted values for X in sample 1 (red crosses).



(b) Re-weighting: We average Y in sample 0 (red circles), but we give more weight to observations whose X is more likely to come from sample 1.

Figure 2. Visual intuition for regression and re-weighting.

In order to answer these questions, we would like to know what the mean of Y would be if we shifted P_X to be as in sample 1, but left $P_{Y|X}$ unchanged as in sample 0, which we denote as:

$$\theta^{\langle 1,0\rangle} = \int y \mathrm{d}P_Y^{\langle 1,0\rangle}(y), \quad \text{for} \quad P_Y^{\langle 1,0\rangle} = \int P_{Y|X}^{(0)} \mathrm{d}P_X^{(1)}.$$

The fundamental challenge is that $P_Y^{\langle 1,0\rangle}$ is a *counterfactual* distribution (Pearl, 2009, §4.5), in the sense that we don't observe data sampled from it, and so we cannot estimate $\theta^{\langle 1,0\rangle}$ directly as a sample average. It is still possible, however, to identify $\theta^{\langle 1,0\rangle}$, as shown in the following lemma:

Lemma 2.1. *Under the regularity conditions given in the appendix, we have the following identification results:*

$$\begin{aligned} \theta^{\langle 1,0\rangle} &= \mathrm{E}_{(1)}[\gamma(X)] && \text{(REG)} \\ &= \mathrm{E}_{(0)}[\alpha(X)Y], && \text{(REW)} \end{aligned}$$

where $\gamma(X) := \mathrm{E}_{(0)}[Y \mid X]$, $\alpha(X) := \mathrm{d}P_X^{(1)}/\mathrm{d}P_X^{(0)}(X)$, *and* $\mathrm{E}_{(t)}[\cdot]$ *denotes the expectation conditional on* $T = t$.

The intuition for Lemma 2.1 is captured graphically in Figure 2. Equation (REG) gives identification by *regression*.

We learn the dependence between Y and X in sample 0 through a (non-parametric) regression, and then average that regression function over the X in sample 1. This is essentially a non-parametric generalization of the Oaxaca-Blinder decomposition (Oaxaca, 1973; Blinder, 1973). It also appears in the causality literature as part of the mediation formula (see Appendix B and Pearl, 2009, §4.5). Equation (REW) gives identification by *re-weighting*. We average Y in sample 0, but we give more weight to observations whose X is more likely to come from sample 1. Formally, the weights $\alpha(X)$ are Radon-Nykodim (RN) derivatives (Billingsley, 1995). When X has a density or a probability mass function, these are simply the ratio of densities or probability mass functions between the two samples, respectively. The re-weighting idea has been used in the literature on covariate shift problems (Shimodaira, 2000; Bickel et al., 2009), but we are not aware of causal change attribution methods that leverage this insight.

Remark 2.2 (Relation to Mediation). Causal Change Attribution is closely related to decomposing the total effect of an intervention that sets $T = 1$ into the Natural Direct Effect and the Natural Indirect Effect of Pearl (2009, §4.5). We discuss this connection in detail in Appendix B. △

The following result combines regression and re-weighting methods to obtain a more robust identifying equation. It is essentially a version without pre-treatment covariates of the efficient influence function given in Tchetgen-Tchetgen & Shpitser (2012) for causal mediation analysis with a single mediator. We restate it here in our notation, because it will make the intuition of our novel result for a general causal graph (Section 2.3) clearer.

Lemma 2.3. *Let $g(X)$, $a(X)$ be two functions such that $\mathrm{E}_{(0)}[g(X)^2] < \infty$, $\mathrm{E}_{(0)}[a(X)^2] < \infty$. Consider the following estimating equation:*

$$\mathrm{E}_{(1)}[g(X)] + \mathrm{E}_{(0)}[a(X)e(X,Y)], \qquad \text{(DR)}$$

where $e(X,Y) := Y - g(X)$.

Under the conditions of Lemma 2.1, (DR) *is equal to $\theta^{\langle 1,0\rangle}$ if $g(X) = \gamma(X)$ or $a(X) = \alpha(X)$, but not necessarily both.*

Equation (DR) is *doubly robust* in the following sense: it still identifies the parameter of interest even if one of the regression function or the weights is misspecified. The term $\mathrm{E}_{(0)}[a(X)e(X,Y)]$ is a "debiasing" term, as in the double/debiased machine learning literature (Chernozhukov et al., 2018a), consisting of an average of the non-parametric regression error $e(Y,X)$ weighted by $a(X)$. If the regression function is correctly specified, this average will be zero regardless of the weights (by the Law of Iterated Expectations). On the other hand, if the regression function is not correctly specified but the weights are, the second term will account and correct for the misspecification of $g(X)$. We refer the reader to the proof in Appendix A for details.

Remark 2.4 (On the Overlap Assumption). One of the regularity conditions (Assumption A.1) imposes that the support of $P_X^{(0)}$ includes the support of $P_X^{(1)}$. From a technical perspective, this assumption guarantees that the RN derivative $\alpha(X)$ exists. In this remark, we give a more intuitive explanation. The regression strategy (REG) requires estimating a regression of $h(Y)$ on X in sample 0, and then obtaining the fitted values in sample 1. Without overlap, we would be extrapolating. In general, we want to avoid this (unless we have a credibly good parametric model for the regression). Similarly, the re-weighting strategy (REW) breaks down without overlap, because some regions of X values that have positive probability in sample 1 are never observed in sample 0. For our asymptotic results in Section 2.5 we will impose a stronger form of overlap (Assumption A.3), which is analogous to the overlap assumption in ATE estimation under unconfoundedness. △

2.3. Identification

We are now ready to introduce the main result. Let $\boldsymbol{c} := \langle c_1, \ldots, c_K, c_{K+1}\rangle \in \{0,1\}^{K+1}$ denote a *change vector*, where $c_k = 1$ if we shift the k-th causal mechanism to be as in sample 1, and $c_k = 0$ otherwise. The last entry c_{K+1} indicates whether we want to shift the conditional distribution of the outcome, $P_{Y|\boldsymbol{X}}$. We will denote by $P_Y^{\boldsymbol{c}}$ the distribution:

$$P_Y^{\boldsymbol{c}} := \int P_{Y|\boldsymbol{X}}^{(c_{K+1})} \prod_{k=1}^{K} \mathrm{d}P_{X_k|\mathrm{PA}_k}^{(c_k)}.$$

For example, in Section 2.2 we considered $P_Y^{\langle 1,0\rangle}$, where we shifted P_X to be as in sample 1 but kept $P_{Y|X}$ to be as in sample 0. Of all the possible $P_Y^{\boldsymbol{c}}$ for $\boldsymbol{c} \in \{0,1\}^{K+1}$, only two are observed directly in the data: $P_Y^{(0)} := P_Y^{\langle 0,\ldots,0,0\rangle}$ and $P_Y^{(1)} := P_Y^{\langle 1,\ldots,1,1\rangle}$; the rest are counterfactual distributions.

We will quantify differences in the distribution of the outcome through the change in some functional (summary measure) of the form $\theta(P_Y) = \int h(y)\mathrm{d}P_Y(y)$ for some $h\colon \mathbb{R} \to \mathbb{R}$. Important examples of such functionals are the mean $h(y) = y$, the second moment $h(y) = y^2$ (which, combined with the mean, allows to obtain the variance), and the CDF at a point u: $h_u(y) = \mathbb{1}\{y \le u\}$ (which can be inverted to obtain quantiles and Wasserstein distances between two distributions). Our main results concern identification, estimation and inference on $\theta^{\boldsymbol{c}} := \theta(P_Y^{\boldsymbol{c}})$ for a counterfactual distribution described by change vector $\boldsymbol{c} \in \{0,1\}^{K+1}$.

To state our main results, we adopt the following notation. We denote with a bar $\bar{\boldsymbol{X}}_k := (X_1, \ldots, X_k)$ for any $k \le K$.

In a slight abuse of notation, we write $\bar{\boldsymbol{X}}_{K+1} = (\boldsymbol{X}, Y)$ and define $g_{K+1}(\bar{\boldsymbol{X}}_{K+1}) := h(Y)$ as a convention to make mathematical expressions more compact.

Theorem 2.5. *Let $g_k(\bar{\boldsymbol{X}}_k)$, $a_k(\bar{\boldsymbol{X}}_k)$ be any candidate functions such that $\mathrm{E}_{(c_{k+1})}[g_k(\bar{\boldsymbol{X}}_k)^2] < \infty$, $\mathrm{E}_{(c_{k+1})}[a_k(\bar{\boldsymbol{X}}_k)^2] < \infty$ for $k = 1, \dots, K$. Consider the following estimating equation:*

$$\mathrm{E}_{(c_1)}[g_1(X_1)] + \sum_{k=1}^{K} \mathrm{E}_{(c_{k+1})}[a_k(\bar{\boldsymbol{X}}_k) e_k(\bar{\boldsymbol{X}}_{k+1})], \quad \text{(MR)}$$

where $e_k(\bar{\boldsymbol{X}}_{k+1}) := g_{k+1}(\bar{\boldsymbol{X}}_{k+1}) - g_k(\bar{\boldsymbol{X}}_k)$ for $k = 1, \dots, K$.

For $k = 1, \dots, K$ define:

$$\gamma_k(\bar{\boldsymbol{X}}_k) = \mathrm{E}_{(c_{k+1})}[g_{k+1}(\bar{\boldsymbol{X}}_{k+1}) \mid \bar{\boldsymbol{X}}_k],$$
$$\alpha_k(\bar{\boldsymbol{X}}_k) = \prod_{j=1}^{k} \frac{\mathrm{d}P^{(c_j)}_{X_j|\bar{\boldsymbol{X}}_{j-1}}}{\mathrm{d}P^{(c_{k+1})}_{X_j|\bar{\boldsymbol{X}}_{j-1}}}(X_j \mid \bar{\boldsymbol{X}}_{j-1}).$$

Under regularity conditions given in the appendix, (MR) *is equal to $\theta^{\boldsymbol{c}}$ if, for every $k = 1, \dots, K$, either $g_k(\bar{\boldsymbol{X}}_k) = \gamma_k(\bar{\boldsymbol{X}}_k)$ or $a_k(\bar{\boldsymbol{X}}_k) = \alpha_k(\bar{\boldsymbol{X}}_k)$, but not necessarily both.*

The intuition for this result is similar to Lemma 2.3. The parameter of interest can be estimated by *regression*, but now we need to use sequentially nested regressions $\gamma_k(\bar{\boldsymbol{X}}_k)$ (i.e. regressions where the outcome itself is a regression function). This fixes the desired distribution for each causal mechanism. Since we are estimating K regression functions, we require K debiasing terms, which appear additively. Each of these debiasing terms includes a "weight" with true value $\alpha_k(\bar{\boldsymbol{X}}_k)$, which is a product of the RN derivatives for the distributions that are shifted with respect to c_{k+1}. Including these K debiasing terms makes equation (MR) *multiply robust*: for each causal mechanism, only the corresponding regression function or the corresponding weights need to be correctly specified, but not necessarily both. We also note that one of our conditions for multiple robustness is that $g_k(\bar{\boldsymbol{X}}_k) = \gamma_k(\bar{\boldsymbol{X}}_k)$ where $\gamma_k(\bar{\boldsymbol{X}}_k) := \mathrm{E}_{(c_{k+1})}[g_{k+1}(\bar{\boldsymbol{X}}_{k+1}) \mid \bar{\boldsymbol{X}}_k]$. This is weaker than setting $\gamma_k(\bar{\boldsymbol{X}}_k) = \mathrm{E}_{(c_{k+1})}[\gamma_{k+1}(\bar{\boldsymbol{X}}_{k+1}) \mid \bar{\boldsymbol{X}}_k]$, because we do not require that g_{k+1} is correctly specified, as long as $a_{k+1}(\bar{\boldsymbol{X}}_{k+1}) = \alpha_{k+1}(\bar{\boldsymbol{X}}_{k+1})$.

Example 2.6. Suppose $K = 2$ and the causal Markov factorization (F) is $P_{Y|X_1,X_2} P_{X_2|X_1} P_{X_1}$, as captured by the DAG in Figure 3. Suppose we are interested in estimating $\theta^{\boldsymbol{c}}$ for $\boldsymbol{c} = \langle 0, 1, 0 \rangle$. The multiply-robust estimating equation in Theorem 2.5 for this case is:

$$\begin{aligned}&\mathrm{E}_{(0)}[g_1(X_1)] \\ &\quad + \mathrm{E}_{(1)}[a_1(X_1)\{g_2(X_1, X_2) - g_1(X_1)\}] \\ &\quad\quad + \mathrm{E}_{(0)}[a_2(X_1, X_2)\{h(Y) - g_2(X_1, X_2)\}],\end{aligned}$$

where the true values of the unknown functions are:

$$\begin{aligned}\gamma_2(X_1, X_2) &= \mathrm{E}_{(0)}[h(Y) \mid X_1, X_2], \\ \gamma_1(X_1) &= \mathrm{E}_{(1)}[g_2(X_1, X_2) \mid X_1], \\ \alpha_2(X_1, X_2) &= \mathrm{d}P^{(1)}_{X_2|X_1} / \mathrm{d}P^{(0)}_{X_2|X_1}(X_2 \mid X_1), \\ \alpha_1(X_1) &= \mathrm{d}P^{(0)}_{X_1} / \mathrm{d}P^{(1)}_{X_1}(X_1). \qquad \triangledown\end{aligned}$$

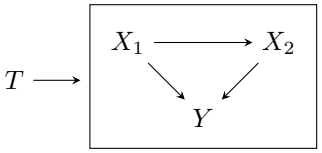


Figure 3. DAG for Example 2.6.

Remark 2.7 (Runtime and Speedup). The general result of Theorem 2.5 implies that, when there are K explanatory variables, we can identify $\theta^{\boldsymbol{c}}$ for any $\boldsymbol{c}$ with a multiply-robust estimating equation that depends on $2K$ unknown functions (K regression functions and K weights) that need to be estimated. For certain change vectors $\boldsymbol{c}$, however, it is possible to simplify the identifying equation to have fewer unknown functions while maintaining multiple robustness. In Appendix C, we discuss two settings where simplifications may occur. The first one is when $c_k = c_{k+1}$ for some k (i.e., when two consecutive regressions are computed with respect to the same probability distribution). The second one is when X_k and X_{k+1} are (conditionally) independent. In particular, when all the explanatory variables are independent, for any given $\boldsymbol{c}$ we only need to estimate one regression and one RN derivative, regardless of K. $\triangle$

2.4. Estimation

In this section, we describe how to implement the multiply-robust estimating equation (MR) in practice.

Step 1. Estimate the Regressions The nested regression functions $\gamma_k(\bar{\boldsymbol{X}}_k)$ can be estimated by any parametric or non-parametric ML regression algorithm, including LASSO, ridge, random forests, neural networks, boosting, or any ensemble method combining multiple of these. We work recursively, starting with an estimator $\hat{\gamma}_K(\bar{\boldsymbol{X}}_K)$ for the regression of $h(Y)$ on $\bar{\boldsymbol{X}}_K := \boldsymbol{X}$, using the the data from sample $c_{K+1} \in \{0, 1\}$ (i.e., the distribution we want to fix for $P_{Y|\boldsymbol{X}}$). Next, we obtain $\hat{\gamma}_{K-1}(\bar{\boldsymbol{X}}_{K-1})$ by regressing $\hat{\gamma}_K(\bar{\boldsymbol{X}}_K)$ on $\bar{\boldsymbol{X}}_{K-1}$, using the the data from sample $c_K \in \{0, 1\}$ (i.e., the distribution we want to fix for $P_{X_K|\mathrm{PA}_K}$). We proceed this way up until we obtain $\hat{\gamma}_1(X_1)$.

Step 2. Estimate the Weights We propose two different approaches to estimate the RN derivatives that appear in the weights $\alpha_k(\bar{\boldsymbol{X}}_k)$: *classification* and *automatic estimation*. Both methods target the weights directly, rather than

estimating a density or probability mass function in each sample and then taking the ratio. For conciseness, we only describe estimation of $\mu(\bar{\boldsymbol{X}}_j) := \mathrm{d}P^{(1)}_{\bar{\boldsymbol{X}}_j}/\mathrm{d}P^{(0)}_{\bar{\boldsymbol{X}}_j}(\bar{\boldsymbol{X}}_j)$. The reciprocal RN derivative can be estimated analogously by exchanging the indices 0 and 1. Conditional RN derivatives for $X_j \mid \bar{\boldsymbol{X}}_{j-1}$ can be obtained by dividing the RN derivative for $\bar{\boldsymbol{X}}_j$ by the RN derivative for $\bar{\boldsymbol{X}}_{j-1}$.

2a. Classification By Bayes' rule, we can express:

$$\mu(\bar{\boldsymbol{X}}_j) = \frac{\beta(\bar{\boldsymbol{X}}_j)}{1-\beta(\bar{\boldsymbol{X}}_j)}\frac{1-p}{p}, \tag{BAY}$$

where $\beta(\bar{\boldsymbol{X}}_j) := \Pr(T = 1 \mid \bar{\boldsymbol{X}}_j)$ and $p := \Pr(T = 1)$ (the posterior and prior probabilities of $\bar{\boldsymbol{X}}_j$ coming from sample 1, respectively). For a given p, this defines a one-to-one mapping between $\mu(\bar{\boldsymbol{X}}_j)$ and $\beta(\bar{\boldsymbol{X}}_j)$.

This suggests the following methodology to estimate $\mu(\bar{\boldsymbol{X}}_j)$. First, we train a classification ML algorithm (logistic regression with LASSO or ridge penalties, neural networks, random forests, etc.) to predict T based on the concatenated data for $\bar{\boldsymbol{X}}_j$. Second, we replace $\beta(\bar{\boldsymbol{X}}_j)$ in Bayes' rule with the predicted posterior probabilities, and we replace $(1-p)/p$ by its empirical analog, n_0/n_1, where n_t is the number of observations in the $T = t$ sample. The posterior probabilities can also be calibrated by cross-validation, as discussed, e.g., in Niculescu-Mizil & Caruana (2005).

Although this method of estimating RN derivatives is not new (c.f., for instance, Sugiyama et al., 2012, pt. II, ch. 4, or Arbour et al., 2021), the application to building a multiply-robust moment function is novel.

2b. Automatic Estimation We specialize general results derived by Chernozhukov et al. (2021) to the causal change attribution problem. In particular, we can characterize the RN derivative $\mu(\bar{\boldsymbol{X}}_j)$ as the solution of:

$$\mu(\bar{\boldsymbol{X}}_j) = \arg\min_m \mathrm{E}_{(0)}[m(\bar{\boldsymbol{X}}_j)^2] - 2\mathrm{E}_{(1)}[m(\bar{\boldsymbol{X}}_j)].$$

The literature refers to this method as "automatic" estimation of the weights because the loss function above depends only in $m(\bar{\boldsymbol{X}}_j)$, without using the explicit form of $\mu(\bar{\boldsymbol{X}}_j)$ as an RN derivative. Thus, we could minimize a sample version of the criterion above over some class of functions (e.g. linear with LASSO or ridge penalties, neural networks, random forests, etc.).

Step 3. Estimate the Target Parameter Finally, we build an estimator of the counterfactual parameter $\theta^{\boldsymbol{c}}$ by replacing $g(\bar{\boldsymbol{X}}_k)$ with $\hat{\gamma}_k(\bar{\boldsymbol{X}}_k)$ and $a(\bar{\boldsymbol{X}}_k)$ with $\hat{\alpha}_k(\bar{\boldsymbol{X}}_k)$ in a sample analog of equation (MR):

$$\hat{\theta}^{\boldsymbol{c}} := \widehat{\mathrm{E}}_{(c_1)}[\hat{\gamma}_1(X_1)] + \sum_{k=1}^{K} \widehat{\mathrm{E}}_{(c_{k+1})}[\hat{\alpha}_k(\bar{\boldsymbol{X}}_k)\hat{e}_k(\bar{\boldsymbol{X}}_{k+1})], \tag{EST}$$

where $\hat{e}_k(\bar{\boldsymbol{X}}_{i,k+1}) := \hat{\gamma}_{k+1}(\bar{\boldsymbol{X}}_{k+1}) - \hat{\gamma}_k(\bar{\boldsymbol{X}}_k)$ for $k = 1, \ldots, K$, and we use $\widehat{\mathrm{E}}_{(t)}[f(W)] = n_t^{-1}\sum_{i:T_i=t} f(W_i)$ to denote the sample average on sample $t \in \{0, 1\}$.

Remark 2.8 (Sample-splitting). To avoid overfitting bias, our recommendation is that the data used to learn the unknown functions $\gamma_k(\bar{\boldsymbol{X}}_k)$, $\alpha_k(\bar{\boldsymbol{X}}_k)$ are not re-used to estimate the main parameter $\theta^{\boldsymbol{c}}$. When the dataset is large, different random subsamples can be used. In medium-sized datasets, a good alternative is to use cross-fitting (see, for example, Newey & Robins, 2018). △

2.5. Large-Sample Inference

In this section, we show consistency and asymptotic normality of our estimator $\hat{\theta}^{\boldsymbol{c}}$, and explain how to perform large-sample inference on the true parameter $\theta^{\boldsymbol{c}}$. First, we notice that the estimator in (EST) is the sum of two averages over different samples:

$$\hat{\theta}^{\boldsymbol{c}} = \widehat{\mathrm{E}}_{(0)}[\hat{\psi}_{(0)}(\bar{\boldsymbol{X}}_{K+1})] + \widehat{\mathrm{E}}_{(1)}[\hat{\psi}_{(1)}(\bar{\boldsymbol{X}}_{K+1})],$$

where $\hat{\psi}_{(t)}(\bar{\boldsymbol{X}}_{K+1})$ contains the terms that depend on the data from sample $t \in \{0, 1\}$ (the exact expression is given in the proof of Theorem 2.9). This characterization will be useful in the results below in terms of writing the asymptotic variance of $\hat{\theta}^{\boldsymbol{c}}$. We define an "oracle" version of the same object, $\psi_{(t)}(\bar{\boldsymbol{X}}_{K+1})$, by replacing $\hat{\gamma}_k(\bar{\boldsymbol{X}}_k), \hat{\alpha}_k(\bar{\boldsymbol{X}}_k)$ with their respective true values $\gamma_k(\bar{\boldsymbol{X}}_k), \hat{\alpha}_k(\bar{\boldsymbol{X}}_k)$ for $k = 1, \ldots, K$.

The following theorem gives an asymptotic normality result for $\hat{\theta}^{\boldsymbol{c}}$, which allow us to perform valid inference on the true parameter $\theta^{\boldsymbol{c}}$ in large samples (e.g., provide standard errors or confidence intervals):

Theorem 2.9. *Under regularity conditions given in the appendix, $\hat{\theta}^{\boldsymbol{c}}$ is consistent and asymptotically normal:*

$$\hat{\theta}^{\boldsymbol{c}} \xrightarrow{p} \theta^{\boldsymbol{c}} \quad \textit{and} \quad \sqrt{n}(\hat{\theta}^{\boldsymbol{c}} - \theta^{\boldsymbol{c}}) \xrightarrow{d} N(0, V),$$

where $n := n_0 + n_1$ and

$$V := \frac{1}{1-p}\mathrm{Var}_{(0)}[\psi_{(0)}(\bar{\boldsymbol{X}}_{K+1})] + \frac{1}{p}\mathrm{Var}_{(1)}[\psi_{(1)}(\bar{\boldsymbol{X}}_{K+1})].$$

Moreover, V can be estimated consistently by:

$$\hat{V} := \frac{n}{n_0}\widehat{\mathrm{Var}}_{(0)}[\hat{\psi}_{(0)}(\bar{\boldsymbol{X}}_{K+1})] + \frac{n}{n_1}\widehat{\mathrm{Var}}_{(1)}[\hat{\psi}_{(1)}(\bar{\boldsymbol{X}}_{K+1})]$$

and $\Pr(\theta^{\boldsymbol{c}} \in [\hat{\theta}^{\boldsymbol{c}} \mp z_{1-a/2} \times (\hat{V}/n)^{1/2}]) \to 1 - a$, where $z_{1-a/2}$ be the $(1-a/2)$-th quantile of a standard Gaussian random variable.

Remark 2.10 (On the Rate Conditions). Assumption A.5, given in the appendix, imposes certain requirements on the root mean-square (RMSE) convergence rates of $\hat{\gamma}_k$ and $\hat{\alpha}_k$, which depend on the choice of algorithm and the properties of the true regression and weighting functions (smoothness, sparsity, etc.). The main condition is that the product of the RMSE for the regression and for the weights has to vanish faster than $n^{-1/2}$. This restriction guarantees that the bias of the main estimator $\hat{\theta}^{\boldsymbol{c}}$ is small enough in large samples to obtain valid inference. The product condition captures an important trade-off: in situations where the regression functions can be estimated at a fast rate, the method will work even if the weights converge very slowly, and vice versa, which is a key advantage of using a multiply-robust estimating equation. △

2.6. Quantifying the Contribution of Each Causal Mechanism

We now discuss two ways to use the $\theta^{\boldsymbol{c}}$ for the problem of causal attribution: *Shapley values* and *along a causal path*.

Shapley Values Shapley values compute the effect of shifting a causal mechanism averaging over each possible combination of which other causal mechanisms to change. Formally, let $\boldsymbol{e}_k$ be a $(K+1)$-vector with a 1 in the k-th entry and 0 everywhere else. The Shapley value associated with $P_{X_k|\mathrm{PA}_k}$ is:

$$\mathrm{SHAP}_k = \sum_{\boldsymbol{c}:c_k=0} \frac{1}{(K+1)\binom{K}{\sum_j c_j}} (\theta^{\boldsymbol{c}+\boldsymbol{e}_k} - \theta^{\boldsymbol{c}}).$$

Along a Causal Path Define, for each $k = 1, \ldots, K+1$, a vector $\boldsymbol{b}_k \in \{0,1\}^{K+1}$ with entries $1, \ldots, k$ equal to 1, and the rest equal to 0. In this method, we define the contribution of the k-th causal mechanism to be:

$$\mathrm{PATH}_k = \theta^{\boldsymbol{b}_k} - \theta^{\boldsymbol{b}_{k-1}}.$$

In other words, we define the contribution of the k-th causal mechanism as the effect of shifting from $P_{X_k|\mathrm{PA}_k}$ fixing its causal antecedents to be as in sample 1, but its causal descendants to be as in sample 0.

How to choose between these two approaches? As discussed in the literature (Budhathoki et al., 2021), the effect of changing the one causal mechanism may depend on which other mechanisms have already been changed; Shapley values take that into account, whereas the method along a causal path does not. Whether this is a relevant concern depends on the application: in some cases, the impact of a given explanatory variable may be approximately the same regardless of the distribution of other explanatory variables. On the other hand, causal attribution along a causal path requires estimating $\theta^{\boldsymbol{c}}$ for only K change vectors $\boldsymbol{c}$ (and, by the simplification in Remark C.1, each requires training only one regressor and one classifier), whereas computing Shapley values requires estimating $\theta^{\boldsymbol{c}}$ for 2^{K+1} change vectors $\boldsymbol{c}$ (each of which could require training up to $2K$ machine learners). Some computationally efficient approximations to SHAP_k exist (e.g, Kolpaczki et al., 2023).

Let $\widehat{\mathrm{SHAP}}_k$ and $\widehat{\mathrm{PATH}}_k$ be estimators of SHAP_k and PATH_k, respectively, that replace the true $\theta^{\boldsymbol{c}}$ with the multiply-robust estimator $\hat{\theta}^{\boldsymbol{c}}$ in (EST). The following corollary gives a useful result for large-sample inference on the causal attribution parameters:

Corollary 2.11. *Under the conditions of Theorem 2.9,* $\widehat{\mathrm{SHAP}}_k$ *and* $\widehat{\mathrm{PATH}}_k$ *are consistent and asymptotically normal, with a consistently estimable asymptotic variance.*

This is a consequence of $\widehat{\mathrm{SHAP}}_k$ and $\widehat{\mathrm{PATH}}_k$ being finite linear combinations of $\hat{\theta}^{\boldsymbol{c}}$. The asymptotic variance could be computed explicitly, although this might be cumbersome, especially for Shapley values, since it requires the $2^{K+1} \times 2^{K+1}$ covariance matrix of $\hat{\theta}^{\boldsymbol{c}}$ for $\boldsymbol{c} \in \{0,1\}^{K+1}$. As an alternative, the multiplier bootstrap of Belloni et al. (2017) can be used. We describe the procedure in Appendix D. An advantage of the multiplier approach is that each bootstrap replication does not require re-training the ML estimators of the regression function and weights.

3. Empirical Results

In this section, we discuss two sets of Monte-Carlo simulation results and a real-data application of our method to understanding the gender wage gap. The code for the simulations and the empirical application is available at `https://github.com/victor5as/mr_causal_attribution`.

3.1. Monte-Carlo Simulations

Design 1 We now present simulation evidence on the performance of our method. Our first simulation design is based on the causal model of Example 2.6, with DAG $X_1 \to X_2$, $X_1 \to Y$ and $X_2 \to Y$. We give details about the data-generating process in Appendix E.

Table 1 presents the results of our simulation (over 1,000 draws) for different choices of learners for the regression and the weights. 1a shows the Mean Absolute Error (MAE) for a scenario where both the regression and the weights are correctly specified parametric learners: OLS with quadratic terms for the regressions and logistic classifier with quadratic terms to build the weights.[2] Comparing our multiply-robust (MR) estimator to using regression or

[2]This is the correct specification for the weights, since the log-likelihood ratio between two Gaussian distributions with possibly different variances is a quadratic polynomial.

re-weighting only, we can see that its performance is statistically indistinguishable from the best of the two methods (regression).

Tables 1b and 1c show what happens when either the weights or the regressions are misspecified, while the specification of the other component of the model is still correct. As a particular form of misspecification, we omit the quadratic terms from either the OLS regression or the logistic classifier. In either case, we observe that MR performs as well as the correctly-specified method, while the misspecified method exhibits much higher errors. This illustrates the multiple robustness property of our estimator. When both the weights and the regression are misspecified parametric models, as in 1d, our theoretical guarantees no longer apply. However, we still show these results for completeness. Although it seems to be the case that the error in Shapley values achieved by the MR methods is about as low as the best of the two other methods, it is significantly larger than the correctly-specified benchmark of Table 1a.

Next, we turn our attention to flexible, non-parametric learners for the unknown regression functions and weights, to mimic a more realistic situation in which we do not want to make strong parametric assumptions about these. First, we consider neural networks (Table 1e). We use `MLPRegressor` and `MLPClassifier` from the Python library `sklearn`. We use the default settings for all hyper-parameters, except that we allow for early stopping and use 3 hidden layers of 100 neurons each. Second, we consider gradient boosting (Table 1f), using `GradientBoostingRegressor` and `GradientBoostingClassifier` from `sklearn` with the default settings. In both cases, we calibrate the predicted probabilities after classification with `CalibratedClassifierCV` also from `sklearn`. The results clearly illustrate Remark 2.10: the bias of the MR non-parametric estimator will, in general, vanish faster than the convergence rates of regression-only or re-weighting-only methods. As such, we see that in general MR has equal or lower error than the best of the other two methods. Moreover, the MAE is often statistically indistinguishable from that of the correctly-specified parametric benchmark 1a.

Design 2 In a second set of experiments, we consider the effect of increasing hte number of explanatory variables K. We consider a line DAG, $X_1 \to X_2 \to \cdots \to X_K \to Y$. We give more details about the data-generating process in Appendix E. We use LASSO to learn the regression functions, and Logistic Regression with ℓ_1 penalty (Logit-LASSO) for the weights. In both cases, the penalty is selected by cross-validation.

Table 2 shows the average (across 500 simulations) of the worst-case (across parameters) absolute error, that is, the average of $\max_{k=1,\ldots,K+1} |\mathrm{PATH}_k - \widehat{\mathrm{PATH}}_k|$. The pattern is almost identical if we look at the Mean Absolute Error instead. The MR estimator performs significantly better than the other two methods, even though the regression and the weights are correctly specified (i.e., they are truly linear/logistic and sparse).

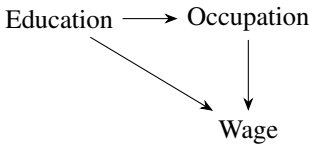


Figure 4. DAG for the gender wage gap application.

3.2. Real-World Application: Gender Wage Gap

Following Budhathoki et al. (2021), we ask how much of the gender wage gap can be attributed to differences in education or occupation. We use data from the Current Population Survey (CPS) 2015. After applying the same sample restrictions as Chernozhukov et al. (2018c), the resulting sample contains 18,137 male and 14,382 female employed individuals. On this data, the (unconditional) gender wage gap in this sample is of \$7.91/hour (standard error 0.01). In other words, female workers earn 24% less on average than male workers; the difference is statistically significant.

We assume the same causal graphical model as Budhathoki et al. (2021), which is captured by the DAG in Figure 4. We randomly split the data into a training set used to estimate the regression function and weights (40%), a calibration set to calibrate the predicted probabilities from our classification algorithm (10%), and an evaluation set used to obtain the main estimates of the $\hat{\theta}^{\boldsymbol{c}}$ parameters (50%). To estimate the regression and the weights, we use `HistGradientBoostingRegressor` and `HistGradientBoostingClassifier` from `sklearn` in Python. We calibrate the probabilities by isotonic regression.

Our estimated Shapley values are plotted in Figure 5. The contribution of education is estimated to be positive and significant. In contrast, the distribution of occupation given education has a significant negative effect, slightly larger in magnitude than the contribution of education, but of opposite sign. Most of the gender wage gap is therefore explained by differences in the distribution of wages given education and occupation. This is consistent with previous findings in the literature. For example, Chernozhukov et al. (2018b), documents the same fact for the UK.

One could think of $P(\text{wage} \mid \text{occup}, \text{educ})$ as "residual" variation, i.e., the remaining gender gap which is not ex-

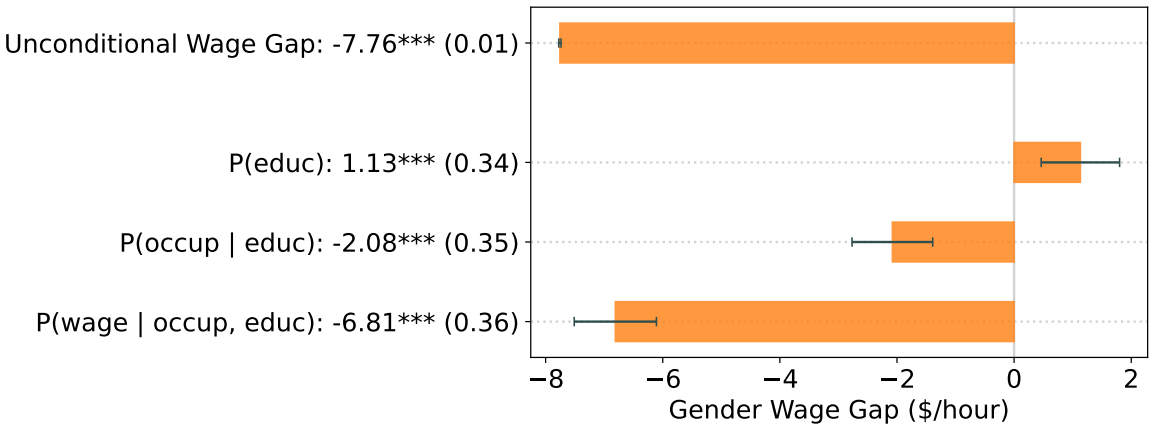


Figure 5. Shapley Values in our gender wage gap application. We plot the point estimates and 95% confidence intervals. We also report the point estimates (bootstrapped standard errors in brackets). We denote statistical significance at the 5% level by **, and at the 1% level by ***.

plained by education or occupation. It may capture the effect of additional variables that are not included in our analysis. For example, there may be gender differences in labor market experience due to women temporarily exiting the labor force to take care of young children. Unfortunately, experience is not available in census data, and so we cannot quantify its impact, but excluding it from the analysis does not bias our results, because it is plausibly not a causal antecedent of education and occupational choice. It may also be due to other factors, including gender discrimination in promotion or gender differences in wage bargaining.

To gain more insight into the results of our attribution exercise, it is natural to ask two questions: (1) how different is the distribution of education and occupation between the two groups, and (2) how do these differences translate into differences in average wage. We present some summary statistics related to these questions in the appendix. Figure 7a compares the distribution of educational attainment for males and females. It appears that females tend to be slightly more educated, with higher rates of college and advanced degree graduation. This could explain the positive Shapley value associated with education; for comparison, college graduates earn $12.60/hour more than high school graduates on average. Figure 7b shows the distribution of occupation conditional on having a college degree. Females are more predominant in administrative, education or healthcare professions, whereas males are more likely to work in management and sales. For comparison, managers earn $16.66/hour more than educators and $6.29/hour more than healthcare practitioners on average. At the same time, there are wide differences across genders that cannot be explained by education and occupation alone: female college graduate managers earn $13.77/hour less on average than their male counterparts, which is consistent with the finding that $P(\text{wage} \mid \text{occup}, \text{educ})$ has the largest Shapley value in our causal change attribution exercise.

4. Conclusions

In this paper, we develop a new estimator for causal change attribution measures, which combines regression and reweighting methods in a multiply-robust estimating equation. We provide the first results on consistency and asymptotic normality of ML-based estimators in this setting, and discuss how to perform inference. Moreover, our method can be used to estimate Shapley values, which will inherit the consistency and large-sample distribution properties.

Finally, we suggest two directions for future research. First, in some settings it will be more plausible to assume unconfoundedness conditional on observed covariates. Tchetgen-Tchetgen & Shpitser (2012) gave a multiply-robust estimator for a single mediator. Extending their results to a general causal graph with multiple explanatory variables would be interesting. Second, the causal interpretation of the change attribution parameters relies on an assumption of no unobserved confounding. The sensitivity bounds in Chernozhukov et al. (2022) could be adapted as a way to test the robustness of a causal change attribution study to unobserved confounding.

Acknowledgements

The authors would like to thank Victor Chernozhukov, Dominik Janzing, Eric Tchetgen-Tchetgen, and seminar audiences at Amazon and MIT for providing helpful feedback. We also would like to thank Patrick Blöbaum for help with code integration in `DoWhy`.

Impact Statement

This paper presents work whose goal is to advance the fields of Causal Inference and Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

The "causal" interpretation of our change attribution measures is grounded on certain assumptions about the Directed Acyclic Graph (DAG) that codifies causal dependence among the variables in the data. We are precise about these assumptions (Section 2.1), and discuss some ways to relax them in future extensions (Section 4). Our theoretical results also rely on regularity conditions, which we state formally in the Appendix and explain in more intuitive terms in the main body of the paper. Practitioners need to be mindful that, if these assumptions do not hold, our proposed algorithm is not guaranteed to lead to accurate results.

References

Arbour, D., Dimmery, D., and Sondhi, A. Permutation weighting. In *International Conference on Machine Learning*, pp. 331–341. PMLR, 2021.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

Berman, R. Beyond the last touch: Attribution in online advertising. *Marketing Science*, 37(5):771–792, 2018.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

Billingsley, P. *Probability and Measure*. John Wiley & Sons, third edition, 1995.

Blinder, A. S. Wage discrimination: reduced form and structural estimates. *Journal of Human resources*, pp. 436–455, 1973.

Blöbaum, P., Götz, P., Budhathoki, K., Mastakouri, A. A., and Janzing, D. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *arXiv preprint arXiv:2206.06821*, 2022.

Budhathoki, K., Janzing, D., Bloebaum, P., and Ng, H. Why did the distribution change? In *AISTATS*, 2021.

Budhathoki, K., Minorics, L., Blöbaum, P., and Janzing, D. Causal structure based root cause analysis of outliers. In *International Conference on Machine Learning*, pp. 2357–2369. PMLR, 2022.

Chen, H., Covert, I. C., Lundberg, S. M., and Lee, S.-I. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, pp. 1–12, 2023.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018a.

Chernozhukov, V., Fernández-Val, I., and Luo, S. Distribution regression with sample selection, with an application to wage decompositions in the UK. Technical Report 1811.11603, arXiv.org, 2018b.

Chernozhukov, V., Fernández-Val, I., and Luo, Y. The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica*, 86(6):1911–1938, 2018c.

Chernozhukov, V., Newey, W. K., Quintas-Martínez, V., and Syrgkanis, V. Automatic debiased machine learning via Riesz regression. *arXiv preprint arXiv:2104.14737*, 2021.

Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A., and Syrgkanis, V. Long Story Short: Omitted Variable Bias in Causal Machine Learning. Technical report, National Bureau of Economic Research, Cambridge, MA, Jul 2022.

Chernozhukov, V., Newey, W., Singh, R., and Syrgkanis, V. Automatic debiased machine learning for dynamic treatment effects and general nested functionals, 2023.

Dalessandro, B., Perlich, C., Stitelman, O., and Provost, F. Causally motivated attribution for online advertising. In *Proceedings of the sixth international workshop on data mining for online advertising and internet economy*, pp. 1–9, 2012.

Daniel, R. M., De Stavola, B. L., Cousens, S. N., and Vansteelandt, S. Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14, 2015.

Dawid, A. P., Faigman, D. L., and Fienberg, S. E. Fitting science into legal contexts: Assessing effects of causes or causes of effects? *Sociological Methods & Research*, 43 (3):359–390, 2014.

Efron, B. Prediction, estimation, and attribution. *International Statistical Review*, 88:S28–S59, 2020.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Janzing, D., Blöbaum, P., Minorics, L., Faller, P., and Mastakouri, A. Quantifying intrinsic causal contributions via structure preserving interventions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024.

Ji, W., Wang, X., and Zhang, D. A probabilistic multi-touch attribution model for online advertising. In *Proceedings of the 25th acm international on conference on information and knowledge management*, pp. 1373–1382, 2016.

Jung, Y., Kasiviswanathan, S., Tian, J., Janzing, D., Blöbaum, P., and Bareinboim, E. On measuring causal contributions via *do*-interventions. In *International Conference on Machine Learning*, pp. 10476–10501. PMLR, 2022.

Kolpaczki, P., Bengs, V., Muschalik, M., and Hüllermeier, E. Approximating the shapley value without marginal contributions. Technical Report 2302.00736, arXiv.org, 2023.

Kulinski, S. and Inouye, D. I. Towards explaining distribution shifts. In *International Conference on Machine Learning*, pp. 17931–17952. PMLR, 2023.

Liu, J., Wang, T., Cui, P., and Namkoong, H. On the need for a language describing distribution shifts: Illustrations on tabular datasets. *arXiv preprint arXiv:2307.05284*, 2023.

Lu, Z., Geng, Z., Li, W., Zhu, S., and Jia, J. Evaluating causes of effects by posterior effects of causes. *Biometrika*, 110(2):449–465, 2023.

Mougan, C., Broelemann, K., Masip, D., Kasneci, G., Thiropanis, T., and Staab, S. Explanation shift: Investigating interactions between models and shifting data distributions. *arXiv preprint arXiv:2303.08081*, 2023.

Newey, W. K. and Robins, J. R. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Niculescu-Mizil, A. and Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005.

Oaxaca, R. Male-female wage differentials in urban labor markets. *International economic review*, pp. 693–709, 1973.

Pearl, J. *Causality*. Cambridge University Press, 2009.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Shao, X. and Li, L. Data-driven multi-touch attribution models. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 258–264, 2011.

Shapley, L. A value for n-person games. Contributions to the theory of games, 1953.

Sharma, A. and Kiciman, E. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

Sharma, A., Li, H., and Jiao, J. The counterfactual-shapley value: Attributing change in system metrics. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, Prediction, and Search*. MIT press, 2000.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Tchetgen-Tchetgen, E. J. and Shpitser, I. Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, 40(3):1816, 2012.

Yamamoto, T. Understanding the past: Statistical analysis of causal attribution. *American Journal of Political Science*, 56(1):237–256, 2012.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 804–813. AUAI, 2011.

Zhao, R., Zhang, L., Zhu, S., Lu, Z., Dong, Z., Zhang, C., Xu, J., Geng, Z., and He, Y. Conditional counterfactual causal effect for individual attribution. In *Uncertainty in Artificial Intelligence*, pp. 2519–2528. PMLR, 2023.

A. Proofs

A.1. Proofs of Section 2.2

Proof of Lemma 2.1. We begin by assuming $P_X^{(1)} \ll P_X^{(0)}$, that is, $P_X^{(1)}$ is absolutely continuous with respect to $P_X^{(0)}$. We give a formal definition of absolute continuity in Assumption A.1 (Weak Overlap), which we will present in the proof of the general identification result in Theorem 2.5. We also discuss this assumption in Remark 2.4 in the main text. By the Radon-Nykodim Theorem (Billingsley, 1995), the Radon-Nykodim (RN) derivative $\alpha(X) := \mathrm{d}P_X^{(1)}/\mathrm{d}P_X^{(0)}(X)$ exists, is almost-surely unique. Furthermore, we assume $\mathrm{E}_{(0)}[Y^2] < \infty$ and $\mathrm{E}_{(0)}[\alpha(X)^2] < \infty$, which is a version of Assumption A.2 (Moments). Assumptions A.1 and A.2, combined with the Cauchy-Schwartz inequality, imply that any function $g(X)$ such that $\mathrm{E}_{(0)}[g(X)^2] < \infty$ will be $P_X^{(1)}$-integrable function $g(X)$, and moreover:

$$\mathrm{E}_{(1)}[g(X)] = \int g(x)\mathrm{d}P_X^{(1)}(x) = \int \alpha(x)g(x)\mathrm{d}P_X^{(0)}(x) = \mathrm{E}_{(0)}[\alpha(X)g(X)] \tag{RN}$$

We're now ready to prove our two identification results. Recall the definition of the target parameter:

$$\theta^{\langle 1,0\rangle} = \int y \,\mathrm{d}P_{Y|X}^{(0)}(y \mid x)\mathrm{d}P_X^{(1)}(x).$$

By A.2, $\gamma(X) := \mathrm{E}_{(0)}[Y \mid X]$ exists and $\mathrm{E}_{(0)}[\gamma(X)^2] < \infty$. By definition of $\gamma(X)$:

$$\mathrm{E}_{(1)}[\gamma(X)] = \int \gamma(x)\mathrm{d}P_X^{(1)}(x) = \int \left\{ \int y \,\mathrm{d}P_{Y|X}^{(0)}(y \mid x) \right\} \mathrm{d}P_X^{(1)}(x) = \int y \,\mathrm{d}P_{Y|X}^{(0)}(y \mid x)\mathrm{d}P_X^{(1)}(x),$$

which shows (REG).

On the other hand, by the Law of Iterated Expectations and property (RN),

$$\mathrm{E}_{(0)}[\alpha(X)Y] = \mathrm{E}_{(0)}[\alpha(X)\gamma(X)] = \mathrm{E}_{(1)}[\gamma(X)],$$

which shows (REW). □

Proof of Lemma 2.3. Consider first the case of $g(X) = \gamma(X)$. In that case, by definition of $\gamma(X)$, $\mathrm{E}[e(Y,X) \mid X] = \mathrm{E}[Y - \gamma(X) \mid X] = 0$. By the Law of Iterated Expectations, $\mathrm{E}_{(0)}[a(X)e(Y,X)] = 0$ for any $a(X)$ such that $\mathrm{E}_{(0)}[a(X)^2] < \infty$, and so (DR) and (REG) give:

$$\mathrm{E}_{(1)}[\gamma(X)] + \mathrm{E}_{(0)}[a(X)e(Y,X)] = \mathrm{E}_{(1)}[\gamma(X)] = \theta^{\langle 1,0\rangle}.$$

If $a(X) = \alpha(X)$, by (RN), $\mathrm{E}_{(1)}[g(X)] = \mathrm{E}_{(0)}[\alpha(X)g(X)]$ for any $g(X)$ such that $\mathrm{E}_{(0)}[g(X)^2] < \infty$, and so (DR) and (REW) give:

$$\mathrm{E}_{(1)}[g(X)] + \mathrm{E}_{(0)}[\alpha(X)Y] - \mathrm{E}_{(0)}[\alpha(X)g(X)] = \mathrm{E}_{(0)}[\alpha(X)Y] = \theta^{\langle 1,0\rangle}.$$ □

A.2. Proofs of Section 2.3

Proof of Theorem 2.5. We assume the following regularity conditions:

Assumption A.1 (Weak Overlap)**.** For all $k = 1, \ldots, K$, $P_{X_k|\mathrm{PA}_k}^{(0)}$ and $P_{X_k|\mathrm{PA}_k}^{(1)}$ are mutually absolutely continuous.

Recall that, for two measures μ, ν on a σ-algebra $\mathcal{B}$, we say that $\nu \ll \mu$ (ν is absolutely continuous with respect to μ) whenever $\mu(B) = 0$ implies $\nu(B) = 0$ for $v \in \mathcal{B}$ (Billingsley, 1995). If $\nu \ll \mu$ and $\mu \ll \nu$, we say that μ and ν are mutually absolutely continuous. Absolute continuity guarantees that the RN derivatives that go into the weight exist. Intuitively, we are requiring that the distribution of $X_k \mid \mathrm{PA}_k$ has the same support in samples 0 and 1. This is essential both for regression methods (avoids extrapolating the regression function) and for re-weighting methods (guarantees that we can find observations that will be similar enough to the other sample), as discussed in Remark 2.4.

Assumption A.2 (Moments)**.** $\mathrm{E}_{(t)}[h(Y)^2] < \infty$ and $\mathrm{E}_{(t)}[\alpha_k(X)^2] < \infty$ for all $k = 1, \ldots, K$ and $t \in \{0, 1\}$.

In particular, Assumption A.2 implies $\mathrm{E}_{(t)}[\gamma_k(Y)^2] < \infty$ for all $k = 1, \ldots, K$ and $t \in \{0, 1\}$.

We can now prove the main result. Recall the definition of the general target parameter:

$$\theta^{\mathbf{c}} = \int h(y) \mathrm{d}P^{(c_{K+1})}_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) \prod_{k=1}^{K} \mathrm{d}P^{(c_k)}_{X_k|\mathrm{PA}_k}(x_k \mid \mathrm{pa}_k).$$

Step 1: Initial Case We begin by showing that, if $g_1(X_1) = \gamma_1(X_1) := \mathrm{E}_{(c_2)}[g_2(X_1, X_2) \mid X_1]$ or $a_1(X_1) = \alpha_1(X_1) := \mathrm{d}P^{(c_1)}_{X_1}/\mathrm{d}P^{(c_2)}_{X_1}(X_1)$ (but not necessarily both), we have:

$$\mathrm{E}_{(c_1)}[g_1(X_1)] + \mathrm{E}_{(c_2)}[a_1(X_1)e_1(X_1, X_2)] = \int g_2(x_1, x_2) \mathrm{d}P^{(c_2)}_{X_2|X_1}(x_2 \mid x_1) \mathrm{d}P^{(c_1)}_{X_1}(x_1).$$

Consider first the case $g_1(X_1) = \gamma_1(X_1)$. Then, $\mathrm{E}_{c_2}[e_1(X_1, X_2) \mid X_1] = 0$ and by the Law of Iterated Expectations $\mathrm{E}_{c_2}[a_1(X_1)e_1(X_1, X_2)] = 0$ for any $a_1(X_1)$ such that $\mathrm{E}_{(c_2)}[a_1(X_1)^2] < \infty$. The definition of $\gamma_1(X_1)$ implies:

$$\mathrm{E}_{(c_1)}[\gamma_1(X_1)] = \int \gamma_1(x_1) \mathrm{d}P^{(1)}_{X_1}(x_1) = \int \left\{ \int g_2(x_1, x_2) \mathrm{d}P^{(c_2)}_{X_2|X_1}(x_2 \mid x_1) \right\} \mathrm{d}P^{(1)}_{X}(x).$$

as we wanted to show.

On the other hand, if $a_1(X_1) = \alpha_1(X_1)$, by (RN), $\mathrm{E}_{(c_1)}[g_1(X_1)] = \mathrm{E}_{(c_2)}[\alpha_1(X_1)g_1(X_1)]$ for any $g_1(X)$ such that $\mathrm{E}_{(c_2)}[g_1(X)^2] < \infty$. Moreover, by the law of iterated expectations and (RN), $\mathrm{E}_{(c_2)}[\alpha_1(X_1)g_2(X_1, X_2)] = \mathrm{E}_{(c_2)}[\alpha_1(X_1)\gamma_1(X_1)] = \mathrm{E}_{(c_1)}[\gamma_1(X_1)]$, so

$$\mathrm{E}_{(c_1)}[g_1(X_1)] + \mathrm{E}_{(c_2)}[\alpha_1(X_1)g_2(X_1, X_2)] - \mathrm{E}_{(c_2)}[\alpha_1(X_1)g_1(X_1)] = \int g_2(x_1, x_2) \mathrm{d}P^{(c_2)}_{X_2|X_1}(x_2 \mid x_1) \mathrm{d}P^{(c_1)}_{X_1}(x_1),$$

as we wanted to show.

Step 2: Induction Suppose that for some $2 \leq j \leq K$, it holds that:

$$\mathrm{E}_{(c_1)}[g_1(X_1)] + \sum_{k=1}^{j-1} \mathrm{E}_{(c_{k+1})}[a_k(\bar{\boldsymbol{X}}_k)e_k(\bar{\boldsymbol{X}}_{k+1})] = \int g_j(\bar{\boldsymbol{X}}_j) \prod_{k=1}^{j} \mathrm{d}P^{(c_k)}_{X_k|\bar{\boldsymbol{X}}_{k-1}}(x_k \mid \bar{\boldsymbol{X}}_{k-1}).$$

We next show that, if

$$g_j(\bar{\boldsymbol{X}}_j) = \gamma_j(\bar{\boldsymbol{X}}_j) := \mathrm{E}_{(c_{j+1})}[g_{j+1}(\bar{\boldsymbol{X}}_{j+1}) \mid \bar{\boldsymbol{X}}_j]$$
$$\text{or} \quad a_j(\bar{\boldsymbol{X}}_j) = \alpha_j(\bar{\boldsymbol{X}}_j) := \prod_{k=1}^{j} \frac{\mathrm{d}P^{(c_k)}_{X_k|\bar{\boldsymbol{X}}_{k-1}}}{\mathrm{d}P^{(c_{j+1})}_{X_k|\bar{\boldsymbol{X}}_{k-1}}}(X_k \mid \bar{\boldsymbol{X}}_{k-1})$$

(but not necessarily both), we have:

$$\mathrm{E}_{(c_1)}[g_1(X_1)] + \sum_{k=1}^{j} \mathrm{E}_{(c_{k+1})}[a_k(\bar{\boldsymbol{X}}_k)e_k(\bar{\boldsymbol{X}}_{k+1})] = \int g_{j+1}(\bar{\boldsymbol{X}}_{j+1}) \prod_{k=1}^{j+1} \mathrm{d}P^{(c_k)}_{X_k|\bar{\boldsymbol{X}}_{k-1}}(x_k \mid \bar{\boldsymbol{X}}_{k-1}).$$

If $g_j(\bar{\boldsymbol{X}}_j) = \gamma_j(\bar{\boldsymbol{X}}_j)$, as before, we have $\mathrm{E}_{c_{j+1}}[e_j(\bar{\boldsymbol{X}}_{j+1}) \mid \bar{\boldsymbol{X}}_j] = 0$, and by the Law of Iterated Expectations $\mathrm{E}_{(c_{j+1})}[a_j(\bar{\boldsymbol{X}}_j)e_j(\bar{\boldsymbol{X}}_{j+1})] = 0$ for any $a_j(\bar{\boldsymbol{X}}_j)$ such that $\mathrm{E}_{(c_{j+1})}[a_j(\bar{\boldsymbol{X}}_j)^2] < \infty$. The inductive hypothesis and the definition of $\gamma_j(\bar{\boldsymbol{X}}_j)$ imply the desired result:

$$\begin{aligned}
&\mathrm{E}_{(c_1)}[g_1(X_1)] + \sum_{k=1}^{j-1} \mathrm{E}_{(c_{k+1})}[a_k(\bar{\boldsymbol{X}}_k)e_k(\bar{\boldsymbol{X}}_{k+1})] + \mathrm{E}_{(c_{j+1})}[a_j(\bar{\boldsymbol{X}}_j)e_j(\bar{\boldsymbol{X}}_{j+1})] \\
&\quad = \int \gamma_j(\bar{\boldsymbol{X}}_j) \prod_{k=1}^{j} \mathrm{d}P^{(c_k)}_{X_k|\bar{\boldsymbol{X}}_{k-1}}(x_k \mid \bar{\boldsymbol{X}}_{k-1}) \\
&\quad = \int \left\{ \int g_{j+1}(\bar{\boldsymbol{X}}_{j+1}) \mathrm{d}P^{(c_{j+1})}_{X_{j+1}|\bar{\boldsymbol{X}}_j}(x_{j+1} \mid \bar{\boldsymbol{X}}_j) \right\} \prod_{k=1}^{j} \mathrm{d}P^{(c_k)}_{X_k|\bar{\boldsymbol{X}}_{k-1}}(x_k \mid \bar{\boldsymbol{X}}_{k-1}). \qquad (*)
\end{aligned}$$

Otherwise, suppose that $a_j(\bar{\boldsymbol{X}}_j) = \alpha_j(\bar{\boldsymbol{X}}_j)$. Notice that $\alpha_j(\bar{\boldsymbol{X}}_j)$ is the RN derivative of

$$\prod_{k=1}^{j} \mathrm{d}P^{(c_k)}_{X_k|\bar{\boldsymbol{X}}_{k-1}} \text{ with respect to } \prod_{k=1}^{j} \mathrm{d}P^{(c_{j+1})}_{X_k|\bar{\boldsymbol{X}}_{k-1}}.$$

The first distribution is a counterfactual distribution if $c_k \neq c_{k'}$ for some $k, k' \leq j$, because some causal mechanisms are distributed as in sample 0, and others as in sample 1. The second distribution, however, is the factual joint distribution of $\bar{\boldsymbol{X}}_j$ in sample c_{j+1}, by the causal Markov assumption and the order of the variable labels. In this light, $\alpha_j(\bar{\boldsymbol{X}}_j)$ satisfies property (RN), and for any $g_j(\bar{\boldsymbol{X}}_j)$ with $\mathrm{E}_{(c_{j+1})}[g_j(\bar{\boldsymbol{X}}_j)^2] < \infty$, we have

$$\mathrm{E}_{(c_{j+1})}[\alpha_j(\bar{\boldsymbol{X}}_j) g_j(\bar{\boldsymbol{X}}_j)] = \int g_j(\bar{\boldsymbol{X}}_j) \prod_{k=1}^{j} \mathrm{d}P^{(c_k)}_{X_k|\bar{\boldsymbol{X}}_{k-1}}(x_k \mid \bar{\boldsymbol{X}}_{k-1}).$$

In particular, by the inductive hypothesis and the Law of Iterated Expectations:

$$\mathrm{E}_{(c_1)}[g_1(X_1)] + \sum_{k=1}^{j-1} \mathrm{E}_{(c_{k+1})}[a_k(\bar{\boldsymbol{X}}_k) e_k(\bar{\boldsymbol{X}}_{k+1})] + \mathrm{E}_{(c_{j+1})}[\alpha_j(\bar{\boldsymbol{X}}_j) g_{j+1}(\bar{\boldsymbol{X}}_{j+1})] - \mathrm{E}_{(c_{j+1})}[\alpha_j(\bar{\boldsymbol{X}}_j) g_j(\bar{\boldsymbol{X}}_j)]$$
$$= \mathrm{E}_{(c_{j+1})}[\alpha_j(\bar{\boldsymbol{X}}_j) \gamma_j(\bar{\boldsymbol{X}}_j)].$$

Finally, property (RN) and $(*)$ give the desired result.

Step 3: Final Step Before we finish, remember that we adopted the notational convention $g_{K+1}(\bar{\boldsymbol{X}}_{K+1}) := h(Y)$. Therefore, the K-th inductive step implies:

$$\mathrm{E}_{(c_1)}[g_1(X_1)] + \sum_{k=1}^{K} \mathrm{E}_{(c_{k+1})}[a_k(\bar{\boldsymbol{X}}_k) e_k(\bar{\boldsymbol{X}}_{k+1})] = \int h(y) \mathrm{d}P^{(c_{K+1})}_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) \prod_{k=1}^{K} \mathrm{d}P^{(c_k)}_{X_k|\bar{\boldsymbol{X}}_{k-1}}(x_k \mid \bar{\boldsymbol{X}}_{k-1}).$$

To conclude the proof, remember that we have labeled the variables so that PA_k are part of $\bar{\boldsymbol{X}}_{k-1}$. Because of the causal Markov factorization assumption, $P_{X_k|\bar{\boldsymbol{X}}_{k-1}} = P_{X_k|\mathrm{PA}_k}$. In other words, once we condition on its direct causes in the DAG, X_k is independent of the rest of variables that are part of $\bar{\boldsymbol{X}}_{k-1}$. □

A.3. Proofs of Section 2.5

Proof of Theorem 2.9. Chernozhukov et al. (2023, Theorem 9) gives an asymptotic normality and inference result for nested regression functionals. Here we discuss the assumptions that allow us to apply their results. In the statement of the assumptions, $C < \infty$ denotes a generic positive and finite constant.

Assumption A.3 (Strong Overlap)**.** There exists $\epsilon > 0$ such that, for all $k = 1, \ldots, K$, $\epsilon \leq \Pr(T = 1 \mid \bar{\boldsymbol{X}}_k) \leq 1 - \epsilon$ almost surely. Moreover, for all $k = 1, \ldots, K$, $\|\hat{\alpha}_k\|_\infty \leq C$.

This assumption requires the "posterior" probability of observation $\boldsymbol{X}_k$ sample 1 to be bounded away from 0 and 1. In other words, it must not be possible to tell which sample any given observation comes from with absolute certainty; thus strengthening Assumption A.1. Notice that, this implies $0 < p < 1$, where $p := \Pr(T = 1)$, and that $\alpha_k(\bar{\boldsymbol{X}}_k)$ will also be almost surely bounded by Bayes' rule (BAY). Since the true weights are bounded, we also impose boundedness of the estimated weights $\hat{\alpha}_k(\bar{\boldsymbol{X}}_k)$.

Assumption A.4 (Higher-order Moments)**.** We assume:

(i) For some $q > 2$ and all $k = 1, \ldots, K$, $\mathrm{E}_{(c_k)}[\gamma_k(\bar{\boldsymbol{X}}_k)^q] \leq C$, $\mathrm{E}_{(c_k)}[\hat{\gamma}_k(\bar{\boldsymbol{X}}_k)^q] \leq C$, $\mathrm{E}_{(c_{k+1})}[\gamma_k(\bar{\boldsymbol{X}}_k)^q] \leq C$, and $\mathrm{E}_{(c_{k+1})}[\hat{\gamma}_k(\bar{\boldsymbol{X}}_k)^q] \leq C$.

(ii) For all $k = 1, \ldots, K$, $\mathrm{E}_{(c_{k+1})}[(\gamma_{k+1}(\bar{\boldsymbol{X}}_{k+1}) - \gamma_k(\bar{\boldsymbol{X}}_k))^2 \mid \bar{\boldsymbol{X}}_k] \leq C$.

(iii) $V := \frac{1}{1-p}\mathrm{Var}_{(0)}[\psi_{(0)}(\bar{\boldsymbol{X}}_{K+1})] + \frac{1}{p}\mathrm{Var}_{(1)}[\psi_{(1)}(\bar{\boldsymbol{X}}_{K+1})] \geq c_0 > 0$.

The first part guarantees that we can apply the Central Limit Theorem. The second part is a regularity condition that allows for heteroskedasticity in each regression, but requires the conditional variance to be bounded. The third part assumes that the asymptotic variance of the target parameter is bounded away from 0. Recall that we defined $\psi_{(t)}(\bar{\boldsymbol{X}}_{K+1})$ to be the part of the estimating equation that depends on data from sample t, namely:

$$\psi_{(t)}(\bar{\boldsymbol{X}}_{K+1}) = \begin{cases} \gamma_1(X_1) + \sum_{k:c_{k+1}=t} \alpha_k(\bar{\boldsymbol{X}}_k) e_k(\bar{\boldsymbol{X}}_{k+1}) & \text{if } c_1 = t, \\ \sum_{k:c_{k+1}=t} \alpha_k(\bar{\boldsymbol{X}}_k) e_k(\bar{\boldsymbol{X}}_{k+1}) & \text{otherwise.} \end{cases}$$

Assumption A.5 (Estimation Rates)**.** Let $\|f\|_{t,2} = (\mathrm{E}_{(t)}[f(X)^2])^{1/2}$ denote the $\mathcal{L}^2(P_X^{(t)})$ norm. For $t \in \{0,1\}$ and all $k = 1, \ldots, K$, we assume:

(i) $\|\hat{\gamma}_k(\bar{\boldsymbol{X}}_k) - \gamma_k(\bar{\boldsymbol{X}}_k)\|_{t,2} = o_p(1)$ and $\|\hat{\alpha}_k(\bar{\boldsymbol{X}}_k) - \alpha_k(\bar{\boldsymbol{X}}_k)\|_{t,2} = o_p(1)$.

(ii) $\sqrt{n}\|\hat{\gamma}_k(\bar{\boldsymbol{X}}_k) - \gamma_k(\bar{\boldsymbol{X}}_k)\|_{t,2}\|\hat{\alpha}_k(\bar{\boldsymbol{X}}_k) - \alpha_k(\bar{\boldsymbol{X}}_k)\|_{t,2} = o_p(1)$ and $\sqrt{n}\|\hat{\gamma}_{k+1}(\bar{\boldsymbol{X}}_{k+1}) - \gamma_{k+1}(\bar{\boldsymbol{X}}_{k+1})\|_{t,2}\|\hat{\alpha}_k(\bar{\boldsymbol{X}}_k) - \alpha_k(\bar{\boldsymbol{X}}_k)\|_{t,2} = o_p(1)$.

Assumption A.5 (i) assumes that both $\hat{\gamma}_k$ and $\hat{\alpha}_k$ are consistent in the mean-square sense for $k = 1, \ldots, K$. Assumption A.5 (ii) imposes a trade-off between the rates of convergence of $\hat{\gamma}_k$ and $\hat{\alpha}_k$ that is often encountered in the double/debiased ML literature (see, for example, Chernozhukov et al., 2018a). As we discuss in the main text (Remark 2.10), the product of the root mean-squared errors of $\hat{\gamma}_k$ and $\hat{\alpha}_k$ has to vanish faster than $n^{-1/2}$. That means that, when we have a high-quality estimator for the regression function, the asymptotic guarantees of the method will still hold even if the weights can only be estimated at a relatively slow rate, and vice versa. □

B. Structural Equations and Potential Outcomes

Suppose we are in the setting of the example in Section 2.2. Here we give a structural equations model (SEM) that implies the DAG in Figure 1a:

$$\begin{aligned} Y &\leftarrow g_Y(T, X, \varepsilon_Y), \\ X &\leftarrow g_X(T, \varepsilon_X), \\ T &\leftarrow \varepsilon_T, \end{aligned}$$

where $(\varepsilon_Y, \varepsilon_X, \varepsilon_T)$ are mutually independent disturbances of arbitrary dimension, and $\leftarrow$ denotes assignment. Shifting the causal mechanism $P_X^{(t)}$ from $t = 0$ to $t = 1$ is equivalent to replacing the structural function $g_X(0, \varepsilon_X)$ with $g_X(1, \varepsilon_X)$. Likewise, shifting the causal mechanism $P_{Y|X}^{(t)}$ from $t = 0$ to $t = 1$ is equivalent to replacing the structural function $g_Y(0, X, \varepsilon_Y)$ with $g_Y(1, X, \varepsilon_Y)$.

An intervention that sets $T = t$, or $do(T = t)$, is associated with *potential outcomes* $X(t) := g_X(t, \varepsilon_X)$ and $Y(t) := g_Y(t, X(t), \varepsilon_X)$ (Imbens & Rubin, 2015). Notice that the SEM does not impose any independence restrictions between $X(0)$, $X(1)$, $Y(0)$, and $Y(1)$. An intervention that sets $T = t$ and $X = x$, or $do(T = t, X = x)$, in turn defines another set of potential outcomes $Y(t, x)$.

Pearl (2009, §4.5) defines the *natural indirect effect* of the transition from $T = 0$ to $T = 1$ as:

$$\mathrm{IE}_{0,1} := \mathrm{E}[Y(0, X(1))] - \mathrm{E}[Y(0)].$$

In words, this is the expected effect of holding $T = 0$ constant, and changing X to whatever value it would have attained if T had been set to 1, $X(1)$. As long as there are no open backdoor paths relative to $T \to X$ and $(T, X) \to Y$,[3]

$$\mathrm{IE}_{0,1} = \theta^{\langle 1,0 \rangle} - \theta^{\langle 0,0 \rangle},$$

where $\theta^{\langle 0,0 \rangle} = \mathrm{E}_{(0)}[Y]$ can be estimated directly by taking the average of Y in sample 0, and $\theta^{\langle 1,0 \rangle}$ can be identified as described in Section 2.2. The mediation formula (Pearl, 2009, equation (4.18)) is capturing the same idea as identification by regression (REG).

[3] More generally, we could allow for pre-treatment covariates W that block those backdoor paths. We discuss this direction for future research in Section 4.

Pearl (2009, §4.5) also defines a *natural direct effect* of the transition from $T = 0$ to $T = 1$ as:

$$\mathrm{DE}_{0,1} := \mathrm{E}[Y(1, X(0))] - \mathrm{E}[Y(0)] = \theta^{\langle 0,1\rangle} - \theta^{\langle 0,0\rangle}.$$

This corresponds to the average effect of an intervention that sets $T = 1$ while, at the same time fixing X to the value it would have attained under $T = 0$.

The definitions above imply two ways of decomposing the total effect of T on Y, as the difference between a direct effect and the indirect effect of the opposite transition:

$$\begin{aligned}\mathrm{E}[Y(1)] - \mathrm{E}[Y(0)] &= (\theta^{\langle 1,1\rangle} - \theta^{\langle 1,0\rangle}) + (\theta^{\langle 1,0\rangle} - \theta^{\langle 0,0\rangle}) = \mathrm{IE}_{0,1} - \mathrm{DE}_{1,0}, \quad \text{but also} \\ &= (\theta^{\langle 1,1\rangle} - \theta^{\langle 0,1\rangle}) + (\theta^{\langle 0,1\rangle} - \theta^{\langle 0,0\rangle}) = \mathrm{DE}_{0,1} - \mathrm{IE}_{1,0}.\end{aligned}$$

Which of the possible decompositions to choose? We suggest two methods in Section 2.6: *Shapley values* (Shapley, 1953; Budhathoki et al., 2021), which average over all possible decompositions, or *along a causal path*, which suggests following an order that respects causal ancestry. We note, however, that our main results are about estimating the $\theta^{\boldsymbol{c}}$ parameters, and so they are valid regardless of the particular decomposition a researcher commits to.

C. Simplifications to the Identifying Equation

In the general formulation of Theorem 2.5, when there are K explanatory variables we need to estimate a total of $2K$ unknown functions (K regressions and K weights). For certain change vectors $\boldsymbol{c}$, however, it is possible to simplify the identifying equation to have fewer unknown functions. In this section, we discuss two settings where simplifications may occur. The first is when $c_k = c_{k+1}$ for some k (i.e., when two consecutive regressions are computed with respect to the same probability distribution), allowing us to apply the Law of Iterated Expectations. The second is when X_k and X_{k+1} are (conditionally) independent.

Remark C.1 (Simplification I: Iterated Expectations). Recall the Law of Iterated expectations:

$$\mathrm{E}[\mathrm{E}[A \mid B, C] \mid B] = \mathrm{E}[A \mid B].$$

We can apply this property when $c_k = c_{k+1}$ for some k (i.e., when two consecutive regressions are computed with respect to the same probability distribution) because:

$$\begin{aligned}\mathrm{E}_{(c_k)}[\gamma_k(\bar{\boldsymbol{X}}_k) \mid \bar{\boldsymbol{X}}_{k-1}] &= \mathrm{E}_{(c_k)}[\mathrm{E}_{(c_{k+1})}[\gamma_{k+1}(\bar{\boldsymbol{X}}_{k+1}) \mid \bar{\boldsymbol{X}}_k] \mid \bar{\boldsymbol{X}}_{k-1}] \\ &= \mathrm{E}_{(c_k)}[\gamma_{k+1}(\bar{\boldsymbol{X}}_{k+1}) \mid \bar{\boldsymbol{X}}_{k-1}].\end{aligned}$$

Thus, we can skip estimation of $\gamma_k(\bar{\boldsymbol{X}}_k)$, and define $\gamma_{k-1}(\bar{\boldsymbol{X}}_{k-1}) := \mathrm{E}_{(c_k)}[g_{k+1}(\bar{\boldsymbol{X}}_{k+1}) \mid \bar{\boldsymbol{X}}_{k-1}]$ directly. Because we are not estimating $\gamma_k(\bar{\boldsymbol{X}}_k)$, we can also drop the debiasing term $\mathrm{E}_{(c_{k+1})}[a_k(\bar{\boldsymbol{X}}_k)e_k(\bar{\boldsymbol{X}}_{k+1})]$, so that we also do not need to estimate the corresponding weight $\alpha_k(\bar{\boldsymbol{X}}_k)$. Finally, we need to adjust the residual in the $(k-1)$-th debiasing term according to the simplification to $e_{k-1}(\bar{\boldsymbol{X}}_{k+1}) = g_{k+1}(\bar{\boldsymbol{X}}_{k+1}) - g_{k-1}(\bar{\boldsymbol{X}}_{k-1})$. △

Remark C.2 (Simplification II: Conditional Independence). Suppose that $A \perp B \mid C$:

$$\mathrm{E}[\mathrm{E}[g(A, B, C) \mid A, C] \mid C] = \mathrm{E}[\mathrm{E}[g(A, B, C) \mid B, C] \mid C],$$

i.e., the order of integration of A and B can be exchanged. We can apply this property when $X_k \perp X_{k+1} \mid \bar{\boldsymbol{X}}_{k-1}$ for some k as:

$$\begin{aligned}\mathrm{E}_{(c_k)}[\gamma_k(\bar{\boldsymbol{X}}_k) \mid \bar{\boldsymbol{X}}_{k-1}] &= \mathrm{E}_{(c_k)}[\mathrm{E}_{(c_{k+1})}[\gamma_{k+1}(\bar{\boldsymbol{X}}_{k+1}) \mid \bar{\boldsymbol{X}}_k] \mid \bar{\boldsymbol{X}}_{k-1}] \\ &= \mathrm{E}_{(c_{k+1})}[\mathrm{E}_{(c_k)}[\gamma_{k+1}(\bar{\boldsymbol{X}}_{k+1}) \mid \bar{\boldsymbol{X}}_{k-1}, X_{k+1}] \mid \bar{\boldsymbol{X}}_{k-1}].\end{aligned}$$

Combined with the previous remark, this re-ordering can reduce the number of functions to estimate, as shown in the following examples. △

Example C.3. Consider $K = 3$ and the causal Markov factorization depicted in Figure 6. (We omit T from the DAG, with the understanding that the distribution of all the variables depends on T.) This factorization implies that $X_2 \perp X_3 \mid X_1$. Suppose we are interested in estimating $\theta^{\boldsymbol{c}}$ for $\boldsymbol{c} = \langle 1, 0, 1, 0\rangle$. Remark C.2 implies that we can exchange the order of

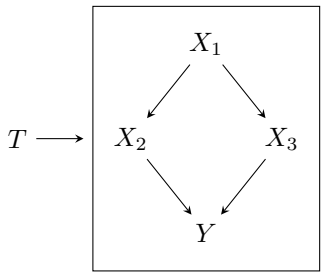


Figure 6. DAG with causal factorization $P^{(t)}_{(\boldsymbol{X},Y)} = P^{(t)}_{Y|X_2,X_3} P^{(t)}_{X_3|X_1} P^{(t)}_{X_2|X_1} P^{(t)}_{X_1}$

integration of X_2 and X_3. Because we're integrating X_1 and X_3 with respect to the same distribution $P^{(1)}_{(\boldsymbol{X},Y)}$, and X_2 and Y with respect to the same distribution $P^{(0)}_{(\boldsymbol{X},Y)}$, Remark C.1 helps us in reducing the number of unknown functions to estimate. In particular, we can use the doubly-robust estimating equation:

$$\mathrm{E}_{(1)}[g_1(X_1, X_3)] + \mathrm{E}_{(0)}[a_1(X_1, X_3)\{h(Y) - g_1(X_1, X_3)\}],$$

where the true value of the unknown functions is:

$$\gamma_1(X_1, X_2) = \mathrm{E}_{(0)}[h(Y) \mid X_1, X_2], \quad \alpha_1(X_1, X_2) = \mathrm{d}P^{(1)}_{X_1,X_3} / \mathrm{d}P^{(0)}_{X_1,X_3}(X_1, X_3).$$

Thus, we only need to estimate 2 unknown functions (rather than 6). ▽

Example C.4. An important example is when all explanatory variables are independent, i.e., the causal Markov factorization is $P_{(\boldsymbol{X},Y)} = P_{Y|\boldsymbol{X}} \prod_{k=1}^{K} P_{X_k}$. For a given change vector $\boldsymbol{c}$, let $\boldsymbol{X}_{\boldsymbol{c}} = (X_k : c_k = 1)$ and $\boldsymbol{X}_{-\boldsymbol{c}} = (X_k : c_k = 0)$. Thanks to Remarks C.1 and C.2 we can give a doubly-robust estimating equation for $\theta^{\boldsymbol{c}}$ that only depends only on 2 unknown functions, regardless of what K is. Whenever $c_{K+1} = 0$ (i.e., we want to keep the conditional distribution of $Y \mid \boldsymbol{X}$ as in sample 0), the doubly-robust estimating equation will be:

$$\mathrm{E}_{(1)}[g_1(\boldsymbol{X}_{\boldsymbol{c}})] + \mathrm{E}_{(0)}[a_1(\boldsymbol{X}_{\boldsymbol{c}})\{h(Y) - g_1(\boldsymbol{X}_{\boldsymbol{c}})\}],$$

where the true value of the unknown functions is:

$$\gamma_1(\boldsymbol{X}_{\boldsymbol{c}}) = \mathrm{E}_{(0)}[h(Y) \mid \boldsymbol{X}_{\boldsymbol{c}}], \quad \alpha_1(\boldsymbol{X}_{\boldsymbol{c}}) = \mathrm{d}P^{(1)}_{\boldsymbol{X}_{\boldsymbol{c}}} / \mathrm{d}P^{(0)}_{\boldsymbol{X}_{\boldsymbol{c}}}(\boldsymbol{X}_{\boldsymbol{c}}).$$

Conversely, if $c_{K+1} = 1$ (i.e., we want to shift the conditional distribution of $Y \mid \boldsymbol{X}$ to be as in sample 1), the doubly-robust estimating equation equation will be:

$$\mathrm{E}_{(0)}[g_1(\boldsymbol{X}_{-\boldsymbol{c}})] + \mathrm{E}_{(1)}[a_1(\boldsymbol{X}_{-\boldsymbol{c}})\{h(Y) - g_1(\boldsymbol{X}_{-\boldsymbol{c}})\}],$$

where the true value of the unknown functions is:

$$\gamma_1(\boldsymbol{X}_{-\boldsymbol{c}}) = \mathrm{E}_{(1)}[h(Y) \mid \boldsymbol{X}_{-\boldsymbol{c}}], \quad \alpha_1(\boldsymbol{X}_{-\boldsymbol{c}}) = \mathrm{d}P^{(0)}_{\boldsymbol{X}_{-\boldsymbol{c}}} / \mathrm{d}P^{(1)}_{\boldsymbol{X}_{-\boldsymbol{c}}}(\boldsymbol{X}_{-\boldsymbol{c}}).$$ ▽

D. Multiplier Bootstrap

In this section, we describe how to apply the multiplier bootstrap procedure of Belloni et al. (2017) to compute standard errors for $\widehat{\mathrm{PATH}}_k$ and $\widehat{\mathrm{SHAP}}_k$.

Let $(\xi^*_i)_{i=1}^n$ be i.i.d. draws, independent of the data, from a random variable with $\mathrm{E}[\xi] = 0$, $\mathrm{E}[\xi^2] = 1$ and $\mathrm{E}[\exp(|\xi|)] < \infty$. Common choices for the distribution of ξ include a re-centered standard exponential distribution (Bayesian bootstrap) and the standard Gaussian distribution (Gaussian multiplier bootstrap). We refer the reader to Belloni et al. (2017) for references.

Define, for each $t \in \{0, 1\}$, $\bar{\psi}_{(t)}(\bar{\boldsymbol{X}}_{K+1}) = \hat{\psi}_{(t)}(\bar{\boldsymbol{X}}_{K+1}) - \widehat{\mathrm{E}}_{(t)}[\hat{\psi}_{(t)}(\bar{\boldsymbol{X}}_{K+1})]$. A multiplier bootstrap draw of $\hat{\theta}^{\boldsymbol{c}}$ can be computed as:

$$\hat{\theta}^{\boldsymbol{c},*} = \hat{\theta}^{\boldsymbol{c}} + \sum_{t \in \{0,1\}} \widehat{\mathrm{E}}_{(t)}[\xi^* \times \bar{\psi}_{(t)}(\bar{\boldsymbol{X}}_{K+1})].$$

For each bootstrap draw of $(\xi_i^*)_{i=1}^n$, the resulting set of $\hat{\theta}^{\mathbf{c},*}$ can be used to compute a bootstrap version of the causal attribution parameters of interest, $\widehat{\mathrm{PATH}}_k^*$ and $\widehat{\mathrm{SHAP}}_k^*$. If we repeat this process many times, the standard deviation of $\widehat{\mathrm{PATH}}_k^*$ and $\widehat{\mathrm{SHAP}}_k^*$ over the bootstrap distribution will be an asymptotically valid estimator for the standard errors of $\widehat{\mathrm{PATH}}_k$ and $\widehat{\mathrm{SHAP}}_k$, respectively.

E. Monte-Carlo Simulations

Design 1 Our simulation design is based on the causal model of Example 2.6. At each simulation draw, we generate two samples of size $n_0 = n_1 = 1000$, according to the following distributions:

$$\begin{aligned} X_1 &\sim N(1, \sigma^2_{(t)}), \\ X_2 \mid X_1 &\sim N(\beta_{(t)} X_1, 1), \\ Y \mid X_1, X_2 &\sim N(X_1 + X_2 + 0.25 X_1^2 + \delta_{(t)} X_2^2, 1), \end{aligned}$$

for $t \in \{0, 1\}$. The parameters $\mu_{(t)} := (\sigma^2_{(t)}, \beta_{(t)}, \delta_{(t)})$ capture the features of the distribution that change between sample 0 and 1, namely the variance of X_1, the dependence between X_2 and X_1, and the dependence between Y and X_2^2. In particular, we choose the following values: $\mu_{(0)} = (1, 0.5, 0.25)$ and $\mu_{(1)} = (1.21, 0.2, -0.25)$. This choice of parameters ensures that the distribution of the simulated data satisfies the regularity conditions in Assumptions A.1 and A.2. The functional of interest will be the mean of Y under different counterfactual distributions. Table 1 presents the results of our simulation (over 1000 draws) for different choices of estimators for the regression and the weights.

Design 2 In a second set of experiments, we consider the effect of increasing K. We consider a line DAG, $X_1 \to X_2 \to \cdots \to X_K \to Y$. In sample $T = 0$, we draw $n_0 = 2,000$ observations according to $X_1 \sim N(0, 1)$ and $X_k \mid X_{k-1} \sim N(0.5 X_{k-1}, 0.75)$ for $k = 2, \ldots, K+1$. For each simulation draw, we randomly select 1/10 of the causal mechanisms to shift their mean by 0.2 in sample $T = 1$, also with $n_1 = 2,000$. In each simulation draw, we compute the $\widehat{\mathrm{PATH}}_k$ attribution measure, using LASSO to estimate the regressions, and Logistic Regression with ℓ_1 penalty (Logit-LASSO) for the weights. In both cases, the penalty is selected by cross-validation. We do this for $K \in \{10, 20, 50, 100\}$.

Table 1. Monte-Carlo simulation results (over 1000 draws). The top panel of each table shows the Mean Absolute Error (MAE) in the estimates of the counterfactual mean parameters $\theta^{\boldsymbol{c}}$ for $\boldsymbol{c} \in \{0,1\}^3$. We omit the results for $\theta^{\langle 0,0,0\rangle}$ and $\theta^{\langle 1,1,1\rangle}$, since these parameters can be estimated as sample means directly from the data. The bottom panel of each table shows the MAE in the estimates of the Shapley values corresponding to each causal mechanism. We compare methods that rely only on regression or re-weighting with our multiply-robust (MR) proposed estimator.

(a) Regression and Weights Correctly Specified

	Mean Absolute Error ± std. err.		
	Regression	Re-weighting	MR
$\langle 0,0,1\rangle$	0.060 ± 0.001	0.065 ± 0.002	0.060 ± 0.001
$\langle 0,1,0\rangle$	0.059 ± 0.001	0.065 ± 0.002	0.059 ± 0.001
$\langle 0,1,1\rangle$	0.057 ± 0.001	0.057 ± 0.001	0.057 ± 0.001
$\langle 1,0,0\rangle$	0.072 ± 0.002	0.075 ± 0.002	0.072 ± 0.002
$\langle 1,0,1\rangle$	0.063 ± 0.002	0.077 ± 0.002	0.063 ± 0.002
$\langle 1,1,0\rangle$	0.064 ± 0.002	0.078 ± 0.002	0.064 ± 0.002
SHAP_1	0.072 ± 0.002	0.073 ± 0.002	0.072 ± 0.002
SHAP_2	0.036 ± 0.001	0.043 ± 0.001	0.037 ± 0.001
SHAP_3	0.040 ± 0.001	0.046 ± 0.001	0.041 ± 0.001

(b) Weights Incorrectly Specified

	Mean Absolute Error ± std. err.		
	Regression	Re-weighting	MR
$\langle 0,0,1\rangle$	0.060 ± 0.001	0.059 ± 0.001	0.060 ± 0.001
$\langle 0,1,0\rangle$	0.059 ± 0.001	0.071 ± 0.002	0.059 ± 0.001
$\langle 0,1,1\rangle$	0.057 ± 0.001	0.074 ± 0.002	0.057 ± 0.001
$\langle 1,0,0\rangle$	0.072 ± 0.002	0.085 ± 0.002	0.072 ± 0.002
$\langle 1,0,1\rangle$	0.063 ± 0.002	0.060 ± 0.001	0.063 ± 0.002
$\langle 1,1,0\rangle$	0.064 ± 0.002	0.101 ± 0.002	0.064 ± 0.002
SHAP_1	0.072 ± 0.002	0.083 ± 0.002	0.072 ± 0.002
SHAP_2	0.036 ± 0.001	0.036 ± 0.001	0.036 ± 0.001
SHAP_3	0.040 ± 0.001	0.064 ± 0.001	0.040 ± 0.001

(c) Regression Incorrectly Specified

	Mean Absolute Error ± std. err.		
	Regression	Re-weighting	MR
$\langle 0,0,1\rangle$	0.130 ± 0.002	0.065 ± 0.002	0.062 ± 0.001
$\langle 0,1,0\rangle$	0.066 ± 0.002	0.065 ± 0.002	0.060 ± 0.001
$\langle 0,1,1\rangle$	0.071 ± 0.002	0.057 ± 0.001	0.057 ± 0.001
$\langle 1,0,0\rangle$	0.089 ± 0.002	0.075 ± 0.002	0.073 ± 0.002
$\langle 1,0,1\rangle$	0.098 ± 0.002	0.077 ± 0.002	0.064 ± 0.002
$\langle 1,1,0\rangle$	0.069 ± 0.002	0.078 ± 0.002	0.066 ± 0.002
SHAP_1	0.084 ± 0.002	0.073 ± 0.002	0.073 ± 0.002
SHAP_2	0.045 ± 0.001	0.043 ± 0.001	0.038 ± 0.001
SHAP_3	0.081 ± 0.002	0.046 ± 0.001	0.042 ± 0.001

(d) Regression and Weights Incorrectly Specified

	Mean Absolute Error ± std. err.		
	Regression	Re-weighting	MR
$\langle 0,0,1\rangle$	0.130 ± 0.002	0.059 ± 0.001	0.113 ± 0.002
$\langle 0,1,0\rangle$	0.066 ± 0.002	0.071 ± 0.002	0.076 ± 0.002
$\langle 0,1,1\rangle$	0.071 ± 0.002	0.074 ± 0.002	0.072 ± 0.002
$\langle 1,0,0\rangle$	0.089 ± 0.002	0.085 ± 0.002	0.089 ± 0.002
$\langle 1,0,1\rangle$	0.098 ± 0.002	0.060 ± 0.001	0.084 ± 0.002
$\langle 1,1,0\rangle$	0.069 ± 0.002	0.101 ± 0.002	0.063 ± 0.001
SHAP_1	0.084 ± 0.002	0.083 ± 0.002	0.084 ± 0.002
SHAP_2	0.045 ± 0.001	0.036 ± 0.001	0.036 ± 0.001
SHAP_3	0.081 ± 0.002	0.064 ± 0.001	0.063 ± 0.001

(e) Non-parametric Regression and Weights (NN)

	Mean Absolute Error ± std. err.		
	Regression	Re-weighting	MR
$\langle 0,0,1\rangle$	0.071 ± 0.002	0.068 ± 0.002	0.061 ± 0.001
$\langle 0,1,0\rangle$	0.079 ± 0.002	0.076 ± 0.002	0.059 ± 0.001
$\langle 0,1,1\rangle$	0.083 ± 0.002	0.059 ± 0.001	0.057 ± 0.001
$\langle 1,0,0\rangle$	0.101 ± 0.003	0.076 ± 0.002	0.073 ± 0.002
$\langle 1,0,1\rangle$	0.080 ± 0.002	0.080 ± 0.002	0.063 ± 0.002
$\langle 1,1,0\rangle$	0.070 ± 0.002	0.097 ± 0.002	0.064 ± 0.002
SHAP_1	0.080 ± 0.002	0.068 ± 0.002	0.072 ± 0.002
SHAP_2	0.053 ± 0.001	0.061 ± 0.001	0.037 ± 0.001
SHAP_3	0.051 ± 0.001	0.061 ± 0.001	0.041 ± 0.001

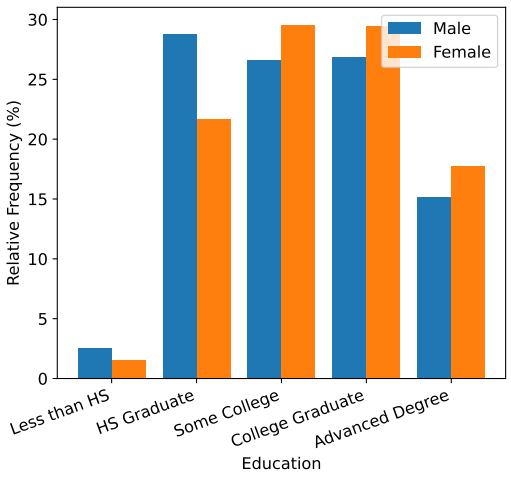
(f) Non-parametric Regression and Weights (GBoost)

	Mean Absolute Error ± std. err.		
	Regression	Re-weighting	MR
$\langle 0,0,1\rangle$	0.060 ± 0.001	0.265 ± 0.002	0.061 ± 0.001
$\langle 0,1,0\rangle$	0.062 ± 0.001	0.080 ± 0.002	0.062 ± 0.001
$\langle 0,1,1\rangle$	0.058 ± 0.001	0.065 ± 0.002	0.058 ± 0.001
$\langle 1,0,0\rangle$	0.074 ± 0.002	0.118 ± 0.002	0.074 ± 0.002
$\langle 1,0,1\rangle$	0.064 ± 0.002	0.280 ± 0.002	0.064 ± 0.002
$\langle 1,1,0\rangle$	0.065 ± 0.002	0.139 ± 0.002	0.065 ± 0.002
SHAP_1	0.072 ± 0.002	0.058 ± 0.001	0.072 ± 0.002
SHAP_2	0.037 ± 0.001	0.120 ± 0.002	0.037 ± 0.001
SHAP_3	0.041 ± 0.001	0.079 ± 0.002	0.041 ± 0.001

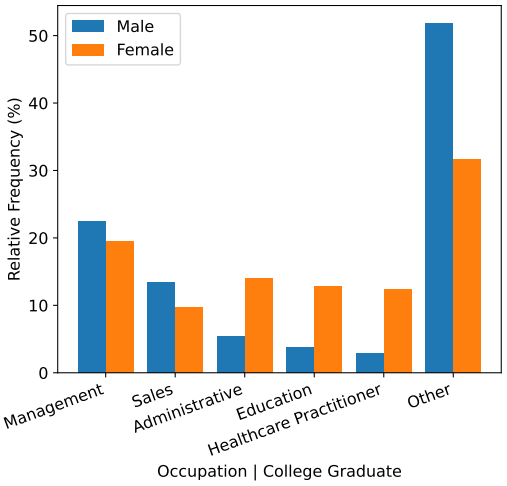
Table 2. Monte-Carlo simulation results (over 5000 draws), Design 2. The table shows the Average Worst-Case Absolute Error (as defined in the main text) for the PATH_k attribution measure.

	Avg. Worst-Case AE $\pm$ std. err.		
K	Regression	Re-weighting	MR
10	0.051 ± 0.001	0.041 ± 0.001	0.030 ± 0.001
20	0.052 ± 0.001	0.042 ± 0.001	0.030 ± 0.001
50	0.050 ± 0.001	0.044 ± 0.001	0.033 ± 0.001
100	0.053 ± 0.001	0.055 ± 0.002	0.044 ± 0.001

F. Additional Figures for the Real-World Application



(a) Female vs. male distribution of education (data from CPS2015).



(b) Female vs. male distribution of occupation conditional on having a college degree (data from CPS2015).

Figure 7. Distribution of explanatory variables by gender in our gender wage gap application.