TECHNICAL REPORT
R-54
November, 2019

Estimating Causal Effects Using Weighting-Based Estimators

Yonghan Jung
Department of Computer Science
Purdue University
jung222@purdue.edu

Jin Tian
Department of Computer Science
Iowa State University
jtian@iastate.edu

Elias Bareinboim
Department of Computer Science
Columbia University
eb@cs.columbia.edu

Abstract

Causal effect identification is one of the most prominent and well-understood problems in causal inference. Despite the generality and power of the results developed so far, there are still challenges in their applicability to practical settings, arguably due to the finitude of the samples. Simply put, there is a gap between causal effect identification and estimation. One popular setting in which sample-efficient estimators from finite samples exist is when the celebrated back-door condition holds. In this paper, we extend weighting-based methods developed for the back-door case to more general settings, and develop novel machinery for estimating causal effects using the weighting-based method as a building block. We derive graphical criteria under which causal effects can be estimated using this new machinery and demonstrate the effectiveness of the proposed method through simulation studies.

1 Introduction

Computing the effects of interventions is one of the central tasks in data-intensive sciences. This problem comes in the literature under the rubric of *causal effect identification* (Pearl 2000, Def. 3.2.4), which asks whether the causal distribution $P(Y = y|do(X = x))$ (for short, $P_x(y)$) can be uniquely identified from a combination of substantive knowledge about the phenomenon under investigation, usually in the form of a causal graph G, and the observational distribution $P(V)$, where V is the set of observed variables. Causal identification has been extensively studied based on the do-calculus (Pearl 1995). Building on this logic, a number of solutions were developed for variants of this problem, including complete graphical and algorithmic conditions (Tian 2002; Huang and Valtorta 2006; Shpitser and Pearl 2006; Bareinboim and Pearl 2012; 2016; Jaber, Zhang, and Bareinboim 2018; Lee, Correa, and Bareinboim 2019).

Even though causal identification has been well-understood and solved in principle, there are still outstanding challenges to the application of these results in practice. By and large, these results assume that the precise observational distribution, $P(V)$, is available for use, while in reality one has access to only a limited number of samples drawn from $P(V)$. One setting where estimators for estimating $P_x(y)$ from finite samples have been systematically developed is when the well-known *back-door (BD)* criterion holds (Pearl 2000, Ch. 3.3.1). That is, if a set of variables Z (called covariates) satisfy the BD criterion relative to (X, Y) then the effect $P_x(y)$ can be identified by covariate adjustment as $P_x(y) = \sum_z P(y|x, z)P(z)$, and the corresponding mean as:

$$\mathbb{E}_{P_x(y)}[Y] = \sum_z \mathbb{E}[Y|x, z]P(z). \qquad (1)$$

Computing Eq. (1) naively – i.e., estimating $\mathbb{E}[Y|x, z]$ and summing over all values $Z = z$ is computationally and statistically challenging whenever the set Z is high dimensional. Regarding the former, summing over $Z = z$ entails an exponential computational burden in $|Z|$, the cardinality of Z; regarding the latter, covering the support of $\mathbb{E}[Y|x, z], P(z)$ with some statistical significance is hardly realizable.

A series of robust and efficient estimators for estimating the BD estimand (Eq. (1)) from finite samples have been developed to circumvent these challenges with great practical success, including propensity score matching (Rosenbaum and Rubin 1983), inverse-probability or stabilized weighting (IPW, SW) (Horvitz and Thompson 1952; Robins, Hernan, and Brumback 2000), doubly robust (Bang and Robins 2005), target maximum likelihood estimator (TMLE) (Van Der Laan and Rubin 2006), and outcome-regression such as BART (Hill 2011), just to cite a few. These techniques have been extended to BD-like estimands for time-series and have been called the g-formula by Robins (1986). This formula holds whenever sequential exchangeability or the sequential back-door (SBD) condition holds (Pearl and Robins 1995).

Despite all their power, these BD-like conditions only cover a limited set of identifiable scenarios, while causal effects could be identifiable in many settings that are not in the form of an adjustment, for which no general purpose estimators have been developed. For instance, we discuss below two practical examples where the causal effects are identifiable but not by BD-like adjustment.

Example 1: Surrogate endpoints. The causal graph in Fig. 1a illustrates a data-generating process of an experimental study that leverages a surrogate endpoint X, a variable intended to substitute for a clinical endpoint Y when the clinical endpoint is hardly accessible. Suppose one is interested in estimating the causal effect of X (e.g., CD4 cell counts) on Y (e.g., Progression of HIV) to validate the CD4 cell counts as a surrogate endpoint (Hughes et al. 1998). W_2 denotes the treatment for the CD4 cell counts and W_1 is a set of confounders affecting the treatment (e.g., a previous disease history). The resultant estimand is given by $P_x(y) = \left(\sum_{w_1} P\left(x, y|w_1, w_2\right) P\left(w_1\right)\right) / \left(\sum_{w_1} P\left(x|w_1, w_2\right) P\left(w_1\right)\right)$, which is clearly not BD-like. To the best of our knowledge, no effective statistical estimator exists for this type of estimands. □

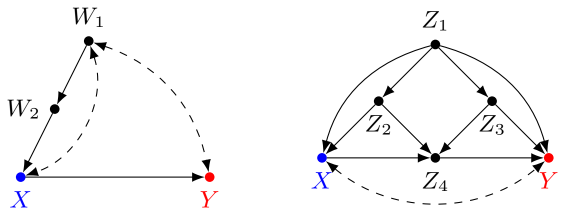
Example 2: Causal mediators. Consider the causal graph in Fig. 1b, where X represents the level of body mass index (BMI), Z_4 the level of multiple, possibly high-dimensional, metabolites, and Y the occurrence of breast cancer (Derkach et al. 2019). Suppose we observe Z_1 (e.g., age), Z_2 (e.g., diets), and Z_3 (e.g., smoking), a set of confounding variables affecting levels of BMI, metabolites, and breast cancer. The goal of the analysis is to assess the effect of BMI levels on breast cancer. The resultant estimand is given by $P_x(y) = \sum_{\mathbf{z}} P(z_4|x, \mathbf{z}^{(3)}) P(\mathbf{z}^{(3)}) \sum_{x'} P(y|x', \mathbf{z}) P(x'|\mathbf{z}^{(3)})$, where $\mathbf{Z} = (Z_1, Z_2, Z_3, Z_4)$ and $\mathbf{Z}^{(3)} = (Z_1, Z_2, Z_3)$, but no statistical estimator is readily available for this estimand. □

In general, many graphical and algorithmic conditions have been developed for determining the identifiability of a causal effect $P_x(y)$ in a given causal graph. However, no general method exists in the literature for estimating $P_x(y)$ from finite samples whenever it is identifiable (for example, as given in Eq. (9)) but not in the form of BD-like adjustment as in Eq. (1)[1]. In short, we note that: given a causal graph G, (i) Complete solutions have been developed for identifying $P_x(y)$ from $P(V)$; (ii) There exist a plethora of methods aiming to estimate BD-like estimands from finite samples when G satisfies the BD/SBD criteria, but the fact is the BD/SBD criteria only capture a small fraction of the scenarios under which causal effects are identifiable; (iii) No systematic treatment exists for estimating arbitrary causal effect estimands that are not BD-like. In this paper, we aim to start bridging the gap between causal "identification" and causal "estimation". Specifically, we propose to extend weighting-based methods developed for BD case (Robins, Hernan, and Brumback 2000) to settings beyond the BD, and further use the weighting-method as a building block to estimate complex causal effect estimands. The contributions of the paper are as follows:

1. We introduce a weighting operator as a building block estimand that could be estimated efficiently using existing statistical techniques developed for the BD estimand.

2. We develop novel machinery for estimating complex causal effects based on the composition of weighting operators.

3. We prove graphical criteria (mSBD, Surrogate, and mSBD composition) that go beyond the BD, under which a causal estimand can be expressed as a weighting operator or their composition, and, therefore, lends itself to effective estimators. Simulation studies demonstrate the effectiveness of the proposed estimators.



(a) Surrogate endpoints

(b) Causal mediators

Figure 1: Causal graphs corresponding to Example 1 and 2. Nodes representing the treatment and outcome are colored in blue and red, respectively.

All the proofs are provided in Appendix D in the supplemental material.

2 Preliminaries

We use the language of structural causal models (SCMs) (Pearl 2000, pp. 204-207) as our basic semantical framework. Each SCM M over a set of variables $\mathbf{V}$ has a causal graph G associated to it. Solid-directed arrows encode functional relationships between observed variables, and dashed-bidirected arrows encode unobserved common causes (e.g., see Fig. 1a). Within the structural semantics, performing an intervention, and setting $\mathbf{X} = \mathbf{x}$, is represented through the do-operator, $do(\mathbf{X} = \mathbf{x})$, which encodes the operation of replacing the original equations of $\mathbf{X}$ by the constant $\mathbf{x}$ and induces a submodel $M_{\mathbf{x}}$ and an experimental distribution $P_{\mathbf{x}}(\mathbf{v})$. Given a causal graph G over a set of variables $\mathbf{V}$, a causal effect $P_{\mathbf{x}}(\mathbf{y})$ is said to be *identifiable* in G if $P_{\mathbf{x}}(\mathbf{y})$ is uniquely computable from $P(\mathbf{v})$ in any SCM that induces G. For a detailed discussion of SCMs, refer to (Pearl 2000).

Each variable will be represented with a capital letter (X) and its realized value with the small letter (x). We will use bold letters ($\mathbf{X}$) to denote sets of variables. Given an ordered set of variables $\mathbf{X} = (X_1, \cdots, X_n)$, we denote $\mathbf{X}^{(i)} = (X_1, \cdots, X_i)$, and $\mathbf{X}^{\geq i} = (X_i, \cdots, X_n)$.

We use the typical graph-theoretic terminology $PA(\mathbf{C}), Ch(\mathbf{C}), De(\mathbf{C}), An(\mathbf{C})$ to represent the union of $\mathbf{C}$ and respectively the parents, children, descendants, and ancestors of $\mathbf{C}$. We use $G_{\overline{\mathbf{C}_1}\underline{\mathbf{C}_2}}$ to denote the graph resulting from deleting all incoming edges to $\mathbf{C}_1$ and all outgoing edges from $\mathbf{C}_2$ in G. $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_G$ denotes that $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{Z}$ in G. $\mathbb{E}[f(\mathbf{Y})|\mathbf{x}]$ denotes the conditional expectation of $f(\mathbf{Y})$ over $P(\mathbf{Y}|\mathbf{x})$. We use $\widehat{P}\left(\mathbf{v}\right)$ to denote the corresponding empirical distribution.

3 Effect Estimation by Weighting Operators

In this section, we start by formally defining a weighting operator as a causal estimand that could be estimated using existing statistical techniques and further used as building blocks to construct more complex causal estimands. We then

[1] Estimators for specific settings, including the SBD and front-door, have been developed based on influence functions (IF) (Fulcher et al. 2019).

present graphical conditions under which a causal estimand can be expressed as a weighting operator.

3.1 Weighting Operator

Causal effect estimation by the BD adjustment is widely used in practice in part due to the availability of efficient estimators from finite samples. In particular, weighting-based statistical estimators for estimating the BD estimand in Eq. (1) have been developed, including the inverse-probability weighting (IPW) and stabilized weighting (SW) (Robins, Hernan, and Brumback 2000). To present weighting techniques, we first define the notion of *weighted distribution* as follows:

Definition 1 (Weighted distribution $P^{\mathcal{W}}(\mathbf{v})$)**.** Given a distribution $P(\mathbf{v})$ and a weight function $\mathcal{W}(\mathbf{v}) > 0$, a weighted distribution $P^{\mathcal{W}}(\mathbf{v})$ is given by

$$P^{\mathcal{W}}(\mathbf{v}) \equiv \frac{\mathcal{W}(\mathbf{v})\, P(\mathbf{v})}{\sum_{\mathbf{v}'} \mathcal{W}(\mathbf{v}')\, P(\mathbf{v}')}. \tag{2}$$

Weighting-based estimators for BD adjustment have been developed based on the following reformulation of the adjustment equation:

Proposition 1. *Let* $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$. *If the causal effect* $P_{\mathbf{x}}(\mathbf{y})$ *is identifiable by the BD adjustment, then* $P_{\mathbf{x}}(\mathbf{y}) = P^{\mathcal{W}}(\mathbf{y}|\mathbf{x})$ *where* $\mathcal{W} = \frac{P(\mathbf{x})}{P(\mathbf{x}|\mathbf{z})}$, *and*

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[\mathbf{Y}] = \mathbb{E}_{P^{\mathcal{W}}(\mathbf{y}|\mathbf{x})}[\mathbf{Y}|\mathbf{X} = \mathbf{x}]. \tag{3}$$

Remarkably, one can estimate $\mathcal{W} = \frac{P(\mathbf{x})}{P(\mathbf{x}|\mathbf{z})}$ as the weight of each individual sample, and treat the reweighted samples as if they were drawn from the causal distribution $P_{\mathbf{x}}(\mathbf{y})$ (Pearl 2000, Ch. 3.6.1). In other words, letting D_{obs} denote samples drawn from $P(\mathbf{x}, \mathbf{y}, \mathbf{z})$, and $D_{obs}^{\mathcal{W}} \sim P^{\mathcal{W}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ represent the reweighted D_{obs}, Prop. 1 says $D_{obs}^{\mathcal{W}}$ plays the role of samples drawn from the post-intervention distribution $P_{\mathbf{x}}(\mathbf{y})$. Therefore, the expected causal effects may be estimated by computing conditional expectation on the reweighted samples. Such weighting-based estimators have also been developed for estimating the g-formula (i.e., g-estimation) (Robins 1986; Robins, Hernan, and Brumback 2000) whenever the SBD condition holds.

In this paper, we will extend the weighting techniques to situations beyond the BD and the g-formula. Towards this goal, we formally define a weighting operator as follows:

Definition 2 (Weighing operator $\mathcal{B}$)**.** Given a weight function $\mathcal{W}(\mathbf{v}) > 0$, a function $h(\mathbf{Y})$, a set of variables $\mathbf{X} = \mathbf{x}$, the weighting operator $\mathcal{B}[h(\mathbf{Y}) \mid \mathbf{x}; \mathcal{W}]$ is defined by

$$\mathcal{B}[h(\mathbf{Y}) \mid \mathbf{x}; \mathcal{W}] \equiv \mathbb{E}_{P^{\mathcal{W}}(\mathbf{y}|\mathbf{x})}[h(\mathbf{Y})|\mathbf{x}] = \sum_{\mathbf{y}} h(\mathbf{y}) P^{\mathcal{W}}(\mathbf{y}|\mathbf{x}).$$

Note that $h(\mathbf{Y})$ is an arbitrary function over $\mathbf{Y}$, and $\mathcal{B}$ is a function of $\mathbf{X} = \mathbf{x}$. We'll describe in Sec. 5 an empirical estimator of the weighting operator $\mathcal{B}$ from finite samples, which extends the existing statistical techniques developed for BD adjustment. Therefore, whenever a causal estimand is expressed as a weighting operator, it will lend itself to effective estimators. In particular, in the form of weighting operator, the BD causal estimand in Prop. 1 is given by $\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[\mathbf{Y}] = \mathcal{B}[\mathbf{Y} \mid \mathbf{x}; \mathcal{W}]$, where $\mathcal{W} = \frac{P(\mathbf{x})}{P(\mathbf{x}|\mathbf{z})}$.

As alluded earlier, the BD-like conditions cover just a limited set of identifiable scenarios. In many settings, causal effects are identifiable but not in the form of an adjustment, and no effective estimators have been developed. In the sequel, we go beyond the BD condition and propose new graphical conditions under which a causal estimand can be expressed as a weighting operator. In Sec. 4, we further show that weighting operators can be used as building blocks to construct more complex causal estimands.

3.2 Multi-outcome Sequential Back-door (mSBD) Criterion and Weighting

One setting of practical interest where the causal estimand can be expressed as a weighting operator is in the time-series domain with a sequence of treatments $X_1, \dots, X_n$ and corresponding covariates $Z_1, \dots, Z_n$. We highlight that the BD criterion has been extended to the sequential BD (SBD) criterion in the time-series domain (Pearl and Robins 1995), where the outcome variable $\mathbf{Y}$ is assumed to be a singleton. Here, we generalize the SBD criterion to accommodate the situation when $\mathbf{Y}$ is a set of variables, for example, for when the outcomes are longitudinal[2].

Definition 3 (Multi-outcome sequential back-door (mSBD) criterion)**.** Given the pair of sets $(\mathbf{X}, \mathbf{Y})$, let $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ be topologically ordered as $X_1 < X_2 < \cdots < X_n$. Let $\mathbf{Y}_0 = \mathbf{Y} \setminus De(\mathbf{X})$ and $\mathbf{Y}_i = \mathbf{Y} \cap \left(De(X_i) \setminus De\left(\mathbf{X}^{\geq i+1}\right)\right)$ for $i = 1, 2, \cdots, n$. Let $ND\left(\mathbf{X}^{\geq i}\right)$ be the set of nondescendants of $\mathbf{X}^{\geq i}$. Then the sequence of variables $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n)$ are said to be mSBD admissible relative to $(\mathbf{X}, \mathbf{Y})$ if it holds that $\mathbf{Z}_i \subseteq ND\left(\mathbf{X}^{\geq i}\right)$, and

$$\left(\mathbf{Y}^{\geq i} \perp\!\!\!\perp X_i | \mathbf{Y}^{(i-1)}, \mathbf{Z}^{(i)}, \mathbf{X}^{(i-1)}\right)_{G_{\underline{X_i}\overline{\mathbf{X}^{\geq i+1}}}}.$$

Roughly speaking, Def. 3 requires that the past observations $\left(\mathbf{X}^{(i-1)}, \mathbf{Y}^{(i-1)}, \mathbf{Z}^{(i)}\right)$ satisfy the BD criterion relative to each $(X_i, \mathbf{Y}^{\geq i})$ pair as covariates. The mSBD criterion reduces to the original SBD (Pearl and Robins 1995) whenever Y is a singleton. When the mSBD criterion holds in a causal graph, the causal effect is identifiable as follows:

Theorem 1 (mSBD adjustment)**.** *If* $\mathbf{Z}$ *is mSBD admissible relative to* $(\mathbf{X}, \mathbf{Y})$, *then* $P_{\mathbf{x}}(\mathbf{y})$ *is identifiable and given by*[3]

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{z}} \prod_{k=0}^{n} P\left(\mathbf{y}_k | \mathbf{x}^{(k)}, \mathbf{z}^{(k)}, \mathbf{y}^{(k-1)}\right) \times \prod_{j=1}^{n} P\left(\mathbf{z}_j | \mathbf{x}^{(j-1)}, \mathbf{z}^{(j-1)}, \mathbf{y}^{(j-1)}\right). \tag{4}$$

[2] Note that treating $\mathbf{Y}$ in SBD criterion as a set would NOT get the mSBD criterion.

[3] We note that the expressions in the form of Eq. (4) or similar are often called the g-formula (Robins, Greenland, and Hu 1999). The mSBD criterion provides a graphical condition under which the causal effect is identifiable as the g-formula.

For example, the causal graph in Fig. 2a represents a time-series setting with a sequence of treatments X_1, X_2, longitudinal outcomes Y_1, Y_2, and corresponding covariates Z_1, Z_2. The BD criterion is not applicable for identifying $P_{x_1,x_2}(y_1, y_2)$. However, (Z_1, Z_2) satisfies the mSBD criterion relative to $((X_1, X_2), (Y_1, Y_2))$. By Thm. 1 $P_{x_1,x_2}(y_1, y_2)$ is identifiable and the expected causal effect of $\{X_1, X_2\}$ on Y_2 is given by

$$\mathbb{E}_{P_{x_1,x_2}(y_2)}[Y_2] = \sum_{z_1,z_2,y_1} \mathbb{E}[Y_2|x_1, x_2, z_1, z_2, y_1]\, P(y_1|x_1, z_1) \times P(z_1)P(z_2|x_1, z_1, y_1) \tag{5}$$

Whenever the mSBD admissible $\mathbf{Z}$ is high-dimensional, evaluating the causal effect is non-trivial in terms of computation and sample efficiency. We address this challenge by leveraging the weighting technique, as shown next.

Theorem 2. *If* $\mathbf{Z}$ *is mSBD admissible relative to* $(\mathbf{X}, \mathbf{Y})$*, then*

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[h(\mathbf{Y})] = \mathcal{B}[h(\mathbf{Y}) \mid \mathbf{x}; \mathcal{W}], \textit{ where} \tag{6}$$

$$\mathcal{W} = \mathcal{W}_{mSBD}(\mathbf{x}, \mathbf{y}, \mathbf{z}) \equiv \frac{P(\mathbf{x})}{\prod_{k=1}^{n} P\left(x_k|\mathbf{x}^{(k-1)}, \mathbf{y}^{(k-1)}, \mathbf{z}^{(k)}\right)}. \tag{7}$$

For example, in Fig. 2a, the expected causal effect of $\{X_1, X_2\}$ on Y_2 can be written, and evaluated, as

$$\mathbb{E}_{P_{x_1,x_2}(y_2)}[Y_2] = \mathcal{B}[Y_2 \mid \{x_1, x_2\}; \mathcal{W}], \tag{8}$$

$$\text{where } \mathcal{W} = \frac{P(x_1, x_2)}{P(x_1|z_1)P(x_2|x_1, y_1, z_1, z_2)}.$$

By Thm. 2, once a set $\mathbf{Z}$ is mSBD-admissible, the expected causal effect can be estimated using the empirical weighting operator described in Sec. 5.

3.3 Surrogate Criterion and Weighting

We present another setting where the causal estimand can be expressed as a weighting operator and can therefore be estimated from finite samples using weighting techniques.

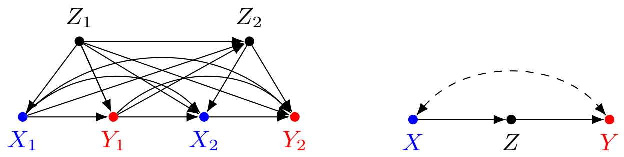
Definition 4 (Surrogate criterion)**.** $(\mathbf{R}, \mathbf{Z})$ is said to be surrogate admissible relative to $(\mathbf{X}, \mathbf{Y})$ if (1) $(\mathbf{Y} \perp\!\!\!\perp \mathbf{R}|\mathbf{X})_{G_{\overline{\mathbf{XR}}}}$; (2) $(\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{R})_{G_{\overline{\mathbf{XR}}}}$; and (3) $\mathbf{Z}$ is mSBD admissible relative to $(\mathbf{R}, (\mathbf{X}, \mathbf{Y}))$.

Theorem 3. *If* $(\mathbf{R}, \mathbf{Z})$ *is surrogate admissible relative to* $(\mathbf{X}, \mathbf{Y})$*, then*[4]

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[h(\mathbf{Y})] = \mathcal{B}[h(\mathbf{Y}) \mid \mathbf{x} \cup \mathbf{r}; \mathcal{W}_{mSBD}(\mathbf{r}, \mathbf{x} \cup \mathbf{y}, \mathbf{z})].$$

To demonstrate the application of the surrogate criterion, we consider Example 1 with its corresponding causal graph given in Fig. 1a, where we are interested in estimating the causal effect of the surrogate endpoint X on the clinical endpoint Y with W_1 being a set of confounders. It can be derived (e.g. by do-calculus) that the causal effect $P_x(y)$ is identifiable and given by

$$P_x(y) = \frac{\sum_{w_1} P(y, x|w_1, w_2)\, P(w_2)}{\sum_{w_1} P(x|w_1, w_2)\, P(w_2)}. \tag{9}$$



(a) mSBD

(b) Front-door

Figure 2: Example graphs

At the first glance, estimating such quotient estimand looks daunting since the variance can be arbitrarily large. To the best of our knowledge, no statistical estimator has been established for the type of estimands like Eq. (9). Thm. 3 provides a solution. By Def. 4, (W_2, W_1) is surrogate admissible relative to (X, Y), and by Thm. 3 we have

$$\mathbb{E}_{P_x(y)}[Y] = \mathcal{B}\left[Y \;\middle|\; \{w_2, x\}; \mathcal{W} = \frac{P(w_2)}{P(w_2|w_1)}\right]. \tag{10}$$

The surrogate criterion allows one to express a complex quotient estimand in the form of a weighting operator, which allows one to estimate through the method discussed in Sec. 5.

4 Causal Effects Estimation by the Composition of Weighting Operators

So far, we have defined a weighting operator as a causal estimand that could be estimated using existing statistical techniques and presented graphical conditions (mSBD and Surrogate criteria) under which a causal estimand can be expressed as a weighting operator. In this section, we introduce novel machinery for causal effect estimation by formulating the front-door estimand as a composition of BD weighting operators. We then extend this idea to develop graphical conditions under which causal effects can be estimated by the composition of weighting operators.

4.1 Estimation of Front-door as a Composition of BD Weighting Operators

A well-known setting where causal effects are identifiable are characterized by what is known as the *front-door* criterion (Pearl 1995), which states that if $\mathbf{Z}$ satisfies the front-door criterion relative to $(\mathbf{X}, \mathbf{Y})$, then the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ is identifiable and is given by the formula

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}) \sum_{\mathbf{x}'} P(\mathbf{y}|\mathbf{x}', \mathbf{z})P(\mathbf{x}'). \tag{11}$$

As an example, consider the causal graph in Fig. 2b, where X represents the level of body mass index (BMI), $\mathbf{Z}$ the level of multiple, possibly high-dimensional, metabolites, and Y the occurrence of breast cancer (Derkach et al. 2019). The goal is to assess the effect of the level of BMI (X) on the breast cancer (Y) in the presence of $\mathbf{Z}$, often called causal mediators. We have that $\mathbf{Z}$ satisfies the front-door criterion relative to (X, Y), and the expected causal effect is given by

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[Y] = \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}) \sum_{\mathbf{x}'} \mathbb{E}[Y|\mathbf{x}', \mathbf{z}]P(\mathbf{x}'). \tag{12}$$

[4] Note the weight function $\mathcal{W}_{\text{mSBD}}$ is defined in Eq. (7).

Computing Eq. (12) is non-trivial in terms of computation and sample efficiency when $\mathbf{Z}$ is high-dimensional. In this paper, we propose a novel method for estimating the front-door estimand. We note something simple albeit powerful: the front-door can be seen as a composition of BD adjustments. To witness, note that:

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{z}} \underbrace{P_{\mathbf{x}}(\mathbf{z})}_{\texttt{BD}=\emptyset} \underbrace{P_{\mathbf{z}}(\mathbf{y})}_{\texttt{BD}=\{\mathbf{X}\}}, \text{ and} \tag{13}$$

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[\mathbf{Y}] = \mathbb{E}_{P_{\mathbf{x}}(\mathbf{z})}\left[\mathbb{E}_{P_{\mathbf{z}}(\mathbf{y})}[\mathbf{Y}]\right], \tag{14}$$

where `BD` represents a BD admissible set, that is, both effects in Eq. (13) can be identified by BD adjustments. In practice, $\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[\mathbf{Y}]$ can be estimated by first estimating $\mathbb{E}_{P_{\mathbf{z}}(\mathbf{y})}[\mathbf{Y}]$, and then estimating the expectation of the resultant quantity over $P_{\mathbf{x}}(\mathbf{z})$, both times using the BD weighting operator. Therefore, we can compute Eq. (12) as a composition of BD weighting operators. Using this example, we formally define a composition of weighting operators as follows:

Definition 5 (Composition of weighting operators)**.** Given two weighting operators $\mathcal{B}_1(\mathbf{x}) \equiv \mathcal{B}[h_z(\mathbf{Z}) \mid \mathbf{x}; \mathcal{W}_1]$ and $\mathcal{B}_2(\mathbf{z}) \equiv \mathcal{B}[h_y(\mathbf{Y}) \mid \mathbf{z}; \mathcal{W}_2]$, the composition of $\mathcal{B}_1$ and $\mathcal{B}_2$ is defined by

$$(\mathcal{B}_1 \circ \mathcal{B}_2)(\mathbf{x}) \equiv \mathcal{B}[\mathcal{B}_2(\mathbf{z}) \mid \mathbf{x}; \mathcal{W}_1]. \tag{15}$$

The front-door estimand (Eq. (12)) can be computed in terms of the composition operation as follows.

Proposition 2. *If $\mathbf{Z}$ satisfies the front-door criterion relative to $(\mathbf{X}, \mathbf{Y})$, then*

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[\mathbf{Y}] = (\mathcal{B}_1 \circ \mathcal{B}_2)(\mathbf{x}), \tag{16}$$

where $\mathcal{B}_1(\mathbf{x}) = \mathcal{B}[h(\mathbf{Z}) \mid \mathbf{x}; \mathcal{W}_1]$, $\mathcal{B}_2(\mathbf{z}) = \mathcal{B}[\mathbf{Y} \mid \mathbf{z}; \mathcal{W}_2]$, $\mathcal{W}_1 = 1$, and $\mathcal{W}_2 = \frac{P(\mathbf{z})}{P(\mathbf{z}|\mathbf{x})}$.

More generally, *we propose using the composition of weighting operators as a novel machinery to construct and estimate complex causal estimands*. The corresponding empirical estimator of the composition of $\mathcal{B}$ operators will be discussed in Sec. 5.

4.2 Causal Effect Estimation by Composition of Weighting Operators

In this section, we study the conditions under which causal effects may be identified by a composition of weighting operators, in which the front-door is just a special case.

Definition 6 (Decomposability criterion)**.** A set of variables $\mathbf{Z}$ satisfies the decomposability criterion relative to $(\mathbf{X}, \mathbf{Y})$ if (1) $(\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{Z})_{G_{\overline{\mathbf{XZ}}}}$; and (2) $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X})_{G_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}}$.

Theorem 4. *If $\mathbf{Z}$ satisfies the decomposability criterion, then*

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{z}} P_{\mathbf{x}}(\mathbf{z}) P_{\mathbf{z}}(\mathbf{y}), \textit{ and}$$

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[h(\mathbf{Y})] = \mathbb{E}_{P_{\mathbf{x}}(\mathbf{z})}\left[\mathbb{E}_{P_{\mathbf{z}}(\mathbf{y})}[h(\mathbf{Y})]\right]. \tag{17}$$

The importance of this theorem lies in that if both causal effects $P_{\mathbf{x}}(\mathbf{z})$ and $P_{\mathbf{z}}(\mathbf{y})$ can be computed using the weighting operators, then $P_{\mathbf{x}}(\mathbf{y})$ can be computed by the composition of weighting operators. In particular, we present a criterion that delineates under what conditions a causal effect can be pieced together through the composition of mSBD weighting operators.

Definition 7 (mSBD composition criterion)**.** Sets of variables $(\mathbf{Z}, \mathbf{W}_1, \mathbf{W}_2)$ are said to satisfy the mSBD composition criterion relative to $(\mathbf{X}, \mathbf{Y})$ if: (1) $\mathbf{Z}$ satisfies the decomposability criterion relative to $(\mathbf{X}, \mathbf{Y})$; and (2) $\mathbf{W}_1$ is mSBD admissible relative to $(\mathbf{X}, \mathbf{Z})$, and $\mathbf{W}_2$ is mSBD admissible relative to $(\mathbf{Z}, \mathbf{Y})$.

Theorem 5 (mSBD composition)**.** *If $(\mathbf{Z}, \mathbf{W}_1, \mathbf{W}_2)$ satisfy the mSBD composition criterion relative to ($\mathbf{X}$, $\mathbf{Y}$), then:*

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[\mathbf{Y}] = (\mathcal{B}_1 \circ \mathcal{B}_2)(\mathbf{x}), \tag{18}$$

where $\mathcal{B}_1(\mathbf{x}) \equiv \mathcal{B}[h(\mathbf{Z}) \mid \mathbf{x}; \mathcal{W}_{mSBD}(\mathbf{x}, \mathbf{z}, \mathbf{w_1})]$ *and* $\mathcal{B}_2(\mathbf{z}) \equiv \mathcal{B}[\mathbf{Y} \mid \mathbf{z}; \mathcal{W}_{mSBD}(\mathbf{z}, \mathbf{y}, \mathbf{w_2})]$.

To demonstrate the application of the mSBD composition criterion, consider the causal mediator scenario (Example 2) with its corresponding causal graph given in Fig. 1b. The set $\mathbf{Z} = (Z_1, Z_2, Z_3, Z_4)$ satisfies the decomposability condition relative to (X, Y), and $(\mathbf{Z}, \emptyset, X)$ satisfy the mSBD composition criterion relative to (X, Y). Therefore, the causal effect $P_x(y)$ can be expressed as $P_x(y) = \sum_{\mathbf{z}} P_x(\mathbf{z}) P_{\mathbf{z}}(y)$. We have that $\emptyset$ satisfies the SBD conditions relative to $(X, (Z_1, Z_2, Z_3, Z_4))$, which yields

$$P_x(\mathbf{z}) = P(z_1, z_2, z_3) P(z_4|z_1, z_2, z_3, x), \tag{19}$$

$$\mathbb{E}_{P_x(\mathbf{z})}[h_{\mathbf{z}}(\mathbf{Z})] = \mathcal{B}[h_{\mathbf{z}}(\mathbf{Z}) \mid \mathbf{x}; \mathcal{W}_1] \equiv \mathcal{B}_1(x), \tag{20}$$

where $\mathcal{W}_1 = \frac{P(x)}{P(x|z_1, z_2, z_3)}$. Further note that $\{X\}$ (i.e. $(\emptyset, \emptyset, \emptyset, X)$) is SBD admissible relative to $((Z_1, Z_2, Z_3, Z_4), Y)$, which yields

$$\mathbb{E}_{P_{\mathbf{z}}(y)}[Y] = \mathcal{B}[Y \mid \mathbf{z}; \mathcal{W}_y] \equiv \mathcal{B}_2(\mathbf{z}), \tag{21}$$

where

$$\mathcal{W}_y = \frac{P(z_1, z_2, z_3, z_4)}{P(z_1, z_2, z_3) P(z_4|z_1, z_2, z_3, x)} = \frac{P(z_4|z_1, z_2, z_3)}{P(z_4|z_1, z_2, z_3, x)}$$

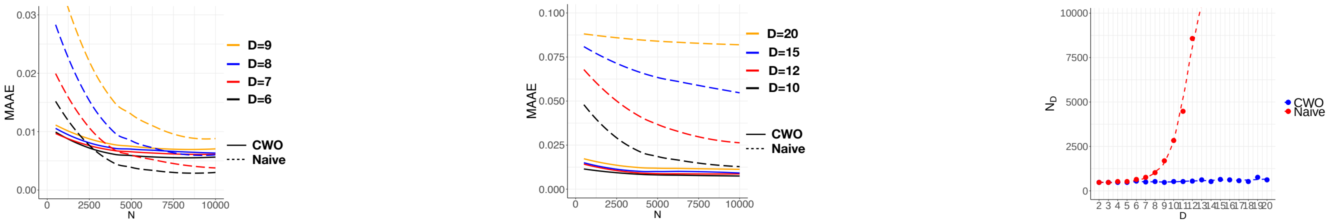
Finally, we obtain that the expected causal effect $\mathbb{E}_{P_x(y)}[Y] = \mathbb{E}_{P_x(\mathbf{z})}\left[\mathbb{E}_{P_{\mathbf{z}}(y)}[Y]\right]$ is given by $(\mathcal{B}_1 \circ \mathcal{B}_2)(x)$.

5 Weighting-based Empirical Estimators

We have introduced the weighting operator as a building block estimand and their composition as a new tool for estimating causal effects. In this section, we present how to estimate the weighting operator and their composition empirically from finite samples. In other words, instead of having access to the true distribution $P(\mathbf{v})$, we only have an i.i.d. data set $D_{obs} = \{\mathbf{V}_{(i)}\}_{i=1}^{N}$ drawn from $P(\mathbf{v})$.

5.1 Empirical Weighting Operators

We extend the weighting-based statistical estimation procedures developed for the BD adjustment to the weighting operator defined in Def. 2. One of the widely used methods



(a) Front-door: low-dim

(b) Front-door: high-dim

(c) D vs. N_D

Figure 3: Experimental results for front-door (Fig. 2b) in which $\mathbf{Z} = (Z_1, \ldots, Z_D)$ consists of D binary variables Z_i: **(a)** MAAE plots with varying $D \in \{6, 7, 8, 9\}$ and **(b)** $D \in \{10, 12, 15, 20\}$; **(c)** The number of samples required to reach predefined estimation error bound D vs. N_D. Plots are best viewed in color.

for estimating the conditional expectation on the weighted samples is the following weighted regression (also known as weighted least square) estimator (Robins, Hernan, and Brumback 2000):

Definition 8 (Empirical weighting operator $\widehat{\mathcal{B}}$)**.** Given $D_{obs} = \{\mathbf{V}_{(i)}\}_{i=1}^{N} \sim P(\mathbf{v})$, the empirical weighting operator $\widehat{\mathcal{B}}\left[h(\mathbf{Y}) \mid \mathbf{x}; \mathcal{W}\right](\mathbf{x}) \equiv g^*(\mathbf{x})$ is estimated by the weighted regression as follows:

$$g^* = \arg\min_{\widehat{g} \in \mathcal{R}} \sum_{i=1}^{N} \widehat{\mathcal{W}}\left(\mathbf{V}_{(i)}\right)\left(h\left(\mathbf{Y}_{(i)}\right) - \widehat{g}\left(\mathbf{X}_{(i)}\right)\right)^2, \qquad (22)$$

where $\widehat{\mathcal{W}}(\mathbf{v})$ is the empirically estimated $\mathcal{W}(\mathbf{v})$, and $\mathcal{R}$ is a class of regression functions (e.g., linear regressions).

For example, for the BD adjustment, we have $\widehat{\mathcal{W}}(\mathbf{V}_{(i)}) = \widehat{P}\left(\mathbf{x}_{(i)}\right) / \widehat{P}\left(\mathbf{x}_{(i)} \mid \mathbf{z}_{(i)}\right)$. When estimating the weight $\widehat{\mathcal{W}}$ from data, in practice, some parametric model will be assumed for $P(\mathbf{x}|\mathbf{z})$, and parameters of the model will be learned from data. When $\mathbf{X} = (X_1, \cdots, X_n)$, one can first use the chain rule of the probability and then model each individual component of $P(\mathbf{x}|\mathbf{z}) = \prod_{d=1}^{n} P\left(x_d|\mathbf{z}, \mathbf{x}^{(d-1)}\right)$. For example, when X is a singleton binary variable, $P(X = 1|\mathbf{z})$ is typically assumed to be a logistic regression function as $(1 + \exp(\alpha_0 + \alpha_{z_1} z_1 + \cdots + \alpha_{z_k} z_k))^{-1}$, and the parameters α are learned from data. Then the trained model is used to estimate the probability. More expressive function classes than logistic regression can be applied to estimate the weights more accurately (Lee, Lessler, and Stuart 2010; Gruber et al. 2015), which may be appealing depending on the particular setting.

Equipped with the estimated weight, one can then estimate the weighting operator by the weighted regression. One can go beyond the standard linear regression class and employ flexible regression functions (Hill 2011; Wen, Hassanpour, and Greiner 2018). We note that $\widehat{\mathcal{B}}$ provides a consistent estimator of $\mathcal{B}$ if the models for $\widehat{\mathcal{W}}$ and $\mathcal{R}$ are correctly specified, following the same argument as in (Robins, Hernan, and Brumback 2000).

Another commonly used method in back-door settings is the Horvitz-Thompson (H-T) estimator (Horvitz and Thompson 1952) as an IPW estimator. We use the weighted regression estimator as the empirical estimator for weighting operators because it has been shown that the H-T estimator may have a higher variance than the weighted regression estimator (Robins, Hernan, and Brumback 2000).

5.2 Estimating Composition of Weighting Operators

Given the empirical weighting operator defined in Def. 8, we simply define the empirical composition of weighting operators as a chain of regressions. Given $\widehat{\mathcal{B}}_1(\mathbf{x}) \equiv \widehat{\mathcal{B}}\left[h_{\mathbf{z}}(\mathbf{Z}) \mid \mathbf{x}; \mathcal{W}_1\right]$ and $\widehat{\mathcal{B}}_2(\mathbf{z}) \equiv \widehat{\mathcal{B}}\left[h_{\mathbf{y}}(\mathbf{Y}) \mid \mathbf{z}; \mathcal{W}_2\right]$, we define $(\widehat{\mathcal{B}_1 \circ \mathcal{B}_2})(\mathbf{x}) \equiv \left(\widehat{\mathcal{B}}_1 \circ \widehat{\mathcal{B}}_2\right)(\mathbf{x})$, which is implemented as $\widehat{\mathcal{B}}\left[\widehat{\mathcal{B}}_2(\mathbf{z}) \,\middle|\, \mathbf{x}; \mathcal{W}_1\right]$, the weighted regression for function $\widehat{\mathcal{B}}_2(\mathbf{z})$ onto $\mathbf{X}$ with weight $\widehat{\mathcal{W}}_1$. Formally,

Definition 9 (Empirical composition of $\mathcal{B}$)**.** Let $\widehat{\mathcal{B}}_1(\mathbf{x}) \equiv \widehat{\mathcal{B}}\left[h_{\mathbf{z}}(\mathbf{Z}) \mid \mathbf{x}; \mathcal{W}_1\right]$ and $\widehat{\mathcal{B}}_2(\mathbf{z}) \equiv \widehat{\mathcal{B}}\left[h_{\mathbf{y}}(\mathbf{Y}) \mid \mathbf{z}; \mathcal{W}_2\right]$. The empirical composition $(\widehat{\mathcal{B}_1 \circ \mathcal{B}_2})(\mathbf{x})$ is defined by

$$(\widehat{\mathcal{B}_1 \circ \mathcal{B}_2})(\mathbf{x}) \equiv \left(\widehat{\mathcal{B}}_1 \circ \widehat{\mathcal{B}}_2\right)(\mathbf{x}) \equiv \widehat{\mathcal{B}}\left[\widehat{\mathcal{B}}_2(\mathbf{z}) \,\middle|\, \mathbf{x}; \mathcal{W}_1\right]. \qquad (23)$$

One question that naturally arises is about the consistency of the empirical composition of weighting operators, which is addressed by the following theorem.

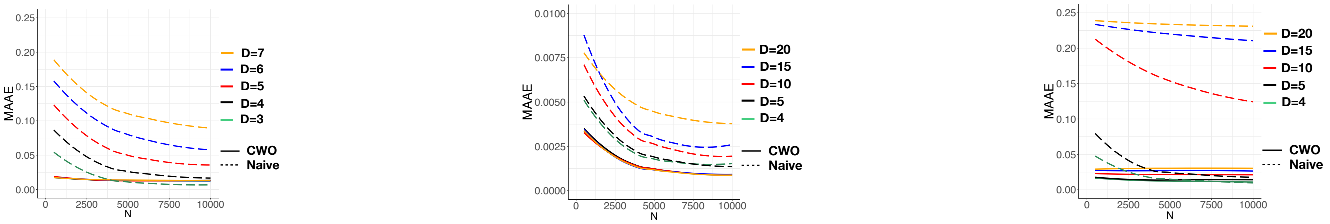
Theorem 6 (Consistency of the composition)**.** *Let $\widehat{\mathcal{B}}_1(\mathbf{x})$ and $\widehat{\mathcal{B}}_2(\mathbf{z})$ be consistent estimators of $\mathcal{B}_1(\mathbf{x})$ and $\mathcal{B}_2(\mathbf{z})$. Let the function class $\mathcal{R}_1$ of $\widehat{\mathcal{B}}_1$ be a compact space. Then, $\left(\widehat{\mathcal{B}}_1 \circ \widehat{\mathcal{B}}_2\right)(\mathbf{x})$ is a consistent estimator of $(\mathcal{B}_1 \circ \mathcal{B}_2)(\mathbf{x})$.*

6 Simulation Studies

6.1 Simulation Setup

Given a causal graph, we will specify a SCM M from which a dataset D_{obs} will be generated. To compute the target $\mu(\mathbf{x}) \equiv \mathbb{E}_{P_{\mathbf{x}}(y)}[Y]$, we generate $N_{int} = 10^7$ number of samples D_{int} from $M_{\mathbf{x}}$, the model from $do(\mathbf{X} = \mathbf{x})$. We estimate $\mu(\mathbf{x})$ by computing the mean of Y in D_{int}, which is treated as the ground truth.

Because there exists no general method in the literature for estimating arbitrary identifiable causal effects that are not in the form of BD-like adjustment, we compare the proposed estimators with a naive procedure, as discussed next:



(a) mSBD (Fig. 2a)

(b) Surrogate endpoints (Fig. 1a)

(c) Causal mediators (Fig. 1b)

Figure 4: MAAE plots for **(a)** mSBD, **(b)** Surrogate endpoints, and **(c)** Causal mediators. Plots are best viewed in color.

Naive procedure As an example, assume we want to evaluate the expression in Eq. (5). We compute each conditional probability such as $P(z_2|x_1, z_1, y_1)$ as $N_{z_2,x_1,z_1,y_1}/N_{x_1,z_1,y_1}$ where $N_{\mathbf{w}}$ is the number of examples in which $\mathbf{W} = \mathbf{w}$. If $N_{x_1,z_1,y_1} = 0$ then $P(z_2|x_1, z_1, y_1)$ is set to zero. $\mathbb{E}\left[Y|x_1, x_2, z_1, z_2, y_1\right]$ is computed as the mean of Y in examples with values $(x_1, x_2, z_1, z_2, y_1)$, and is set to zero if no example has these values. The expected causal effect is computed by summing over all the possible values of Z_1, Y_1, Z_2.

Proposed procedure (named CWO - Composition of Weighting Operators) We use the empirical estimators described in Sec. 5. The conditional probabilities in the weights are estimated by the logistic regression model (binary variables are used in the simulation studies).

Accuracy Measure Given a data set D_{obs} with N examples, let $\mu_{cwo}(\mathbf{x})$ and $\mu_{nai}(\mathbf{x})$ be the estimated $\mathbb{E}_{P_{\mathbf{x}}(y)}\left[Y\right]$ using the CWO and naive procedure respectively. We compute the average absolute error AAE as $|\mu(\mathbf{x}) - \mu_{cwo}(\mathbf{x})|$ averaged over $\mathbf{x}$ and $|\mu(\mathbf{x}) - \mu_{nai}(\mathbf{x})|$ averaged over $\mathbf{x}$ respectively. For each sample size N, we generate 100 data sets. We call the median of the 100 AAEs the *median average absolute error or MAAE*. A plot of MAAE vs. the sample size N will be called a *MAAE plot*.

6.2 Simulation Results

We test the proposed CWO against the naive approach in several scenarios (we only compare with the naive method due to the nonexistence of other general purpose methods applicable in these cases). The detailed descriptions of the corresponding SCMs are provided in Appendix E.

Front-door (Fig. 2b) We first test on the front-door graph for estimating $\mathbb{E}_{P_x(y)}\left[Y\right]$ in Eq. (12). We set X to be binary, Y continuous within $[0, 1]$, and $\mathbf{Z} = (Z_1, \ldots, Z_D)$ with Z_i all binary. Fig. 3a shows MAAE of CWO vs. naive for $D \in \{6, 7, 8, 9\}$, and Fig. 3b the plots for $D \in \{10, 12, 15, 20\}$. We observe that the naive approach works well when $\mathbf{Z}$ is low dimensional (up to $D = 8$) and given many examples. CWO may have bias due to the use of logistic regression models. When $\mathbf{Z}$ is high-dimensional, CWO significantly outperforms the naive approach. To get a better understanding of the sample efficiency, for each given D, we gradually increase the sample size $N = 500, 1000, 1500, \ldots$, and find the corresponding MAAE, and stop to record the sample size N_D when the MAAE is within a predetermined threshold. The threshold was set to 0.025 in these experiments. Roughly, N_D represents how many samples are needed for the estimator to reach a predetermined accuracy. Fig. 3c shows the curves of D vs. N_D. We note that the number of samples needed to reach a predetermined accuracy increases very rapidly (exponentially in D) for the naive approach while CWO scales very well.

mSBD: (Fig. 2a) We test on estimating $\mathbb{E}_{P_{x_1,x_2}(y_2)}\left[Y_2\right]$ given in Eq. (5). We set X_1, X_2, Y_1 to be binary, Y_2 continuous within $[0, 1]$, and $Z_i = (Z_{i1}, \ldots, Z_{iD})$ for $i = 1, 2$, where all Z_{ij} are binary. Fig. 4a presents the MAAE plots for $D \in \{3, 4, 5, 6, 7\}$. We note that CWO provides more robust estimates and significantly outperforms the naive procedure in high-dimensional settings.

Surrogate endpoints (Fig. 1a) We test on estimating $\mathbb{E}_{P_x(y)}\left[Y\right]$ (where the causal effect $P_x(y)$ is given in Eq. (9)). The MAAE plots for $D \in \{4, 5, 10, 15, 20\}$ are given in Fig. 4b. We observe that the CWO method significantly outperforms the naive approach.

Causal mediators (Fig. 1b) We test on estimating $\mathbb{E}_{P_x(y)}\left[Y\right]$. Fig. 4c presents the MAAE plots for $D \in \{4, 5, 10, 15, 20\}$. Again, we note CWO significantly outperforms the naive procedure in high-dimensional settings.

These experimental results show that CWO significantly outperforms its naive counterpart. In Appendix B, we provide a discussion on why CWO outperforms the naive procedure. To better understand to what extent the performance gains over the naive procedure should be attributed to the use of parametric assumptions, we also performed simulations comparing CWO against the parametric plug-in estimator given in Appendix C. Finally, we performed simulations comparing CWO with the H-T estimator given in Appendix G.

7 Conclusions

The problem of determining whether a causal effect is identifiable from observational data given a causal graph is well-understood, while there's virtually no work on how, in general, one can efficiently estimate, from finite samples, an identifiable causal effect beyond BD-like settings. This paper takes the first step in filling in the gap between identification and estimation by developing novel machinery for estimating causal effects through the weighting operators and their composition. We introduced graphical criteria for determining when the new estimation methods are applicable. These results offer new tools for data scientists to be able to estimate effects that the usual methods (including Propensity score, IPW, BART) are not applicable given that the causal estimand is not BD-like. This work opens new research directions. On the one hand, many techniques have been developed for and besides weighted regression for BD estimation; can those techniques be applied and leveraged to the composition of weighting operators? How model misspecification, which is well-studied through double robust methods in the BD-case, should be addressed in this more general setting? On the other hand, can weighting operators be further composed to identify effects beyond the decomposability criterion? Also, can the weighting operator be combined in alternative ways to identify new effects?

Acknowledgements

We thank Sanghack Lee, Daniel Kumor, and the reviewers for all the feedback provided. Elias Bareinboim and Yonghan Jung were supported in parts by grants from NSF IIS-1704352, IIS-1750807 (CAREER), IBM Research, and Adobe Research. Jin Tian was partially supported by NSF grant IIS-1704352 and ONR grant N000141712140.

References

Bang, H., and Robins, J. M. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4):962–973.

Bareinboim, E., and Pearl, J. 2012. Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 113–120. AUAI Press.

Bareinboim, E., and Pearl, J. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* 113(27):7345–7352.

Derkach, A.; Pfeiffer, R. M.; Chen, T.-H.; and Sampson, J. N. 2019. High dimensional mediation analysis with latent variables. *Biometrics*.

Fulcher, I. R.; Shpitser, I.; Marealle, S.; and Tchetgen, E. J. T. 2019. Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Gruber, S.; Logan, R. W.; Jarrín, I.; Monge, S.; and Hernán, M. A. 2015. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Statistics in medicine* 34(1):106–117.

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

Horvitz, D. G., and Thompson, D. J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* 47(260):663–685.

Huang, Y., and Valtorta, M. 2006. Pearl's calculus of intervention is complete. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, 217–224. AUAI Press.

Hughes, M. D.; Daniels, M. J.; Fischl, M. A.; Kim, S.; and Schooley, R. T. 1998. Cd4 cell count as a surrogate endpoint in hiv clinical trials: a meta-analysis of studies of the aids clinical trials group. *Aids* 12(14):1823–1832.

Jaber, A.; Zhang, J.; and Bareinboim, E. 2018. Causal identification under markov equivalence. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

Lee, S.; Correa, J. D.; and Bareinboim, E. 2019. General identifiability with arbitrary surrogate experiments. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

Lee, B. K.; Lessler, J.; and Stuart, E. A. 2010. Improving propensity score weighting using machine learning. *Statistics in medicine* 29(3):337–346.

Pearl, J., and Robins, J. 1995. Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, 444–453. Morgan Kaufmann Publishers Inc.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika* 82(4):669–710.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press. 2nd edition, 2009.

Robins, J. M.; Greenland, S.; and Hu, F.-C. 1999. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association* 94(447):687–700.

Robins, J. M.; Hernan, M. A.; and Brumback, B. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* 11(5).

Robins, J. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling* 7(9-12):1393–1512.

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

Shpitser, I., and Pearl, J. 2006. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence*, 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Tian, J. 2002. *Studies in Causal Reasoning and Learning*. Ph.D. Dissertation, Computer Science Department, University of California, Los Angeles, CA.

Van Der Laan, M. J., and Rubin, D. 2006. Targeted maximum likelihood learning. *The International Journal of Biostatistics* 2(1).

Wen, J.; Hassanpour, N.; and Greiner, R. 2018. Weighted gaussian process for estimating treatment effect. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*.

Appendix
Estimating Causal Effects Using Weighting-Based Estimators

A More Examples on Applications of Weighting Operators and Their Composition

Surrogate-endpoints with mSBD adjustment Fig. 1a (Tian 2002) in Appendix provides more complicated example where there is an ordered set of surrogate experiments (W_2, W_4) on the endpoints (X, Y) and an ordered covariates (W_1, W_3) satisfies the mSBD criterion relative to $((W_2, W_4), (X, Y))$. As a ground truth, we note that the causal effect $P_x(y)$ is given by

$$P_x\left(y\right) = \frac{\sum_{w_1,w_3} P\left(y, x|w_4, w_3, w_2, w_1\right) P\left(w_3|w_2, w_1\right) P\left(w_1\right)}{\sum_{w_1,w_3} P\left(x|w_4, w_3, w_2, w_1\right) P\left(w_3|w_2, w_1\right) P\left(w_1\right)} \tag{1}$$

Since $((W_2, W_4), (W_1, W_3))$ is surrogate admissible relative to (X, Y), the weighting operator encodes the causal effect as follow:

$$\mathbb{E}_{P_x(y)}\left[Y\right] = \mathcal{B}\left[Y \,\middle|\, (x, w_2, w_4)\,; \mathcal{W} = \frac{P\left(w_2, w_4\right)}{P\left(w_2|w_1\right) P\left(w_4|w_3\right)}\right]$$

This exemplifies that complex causal estimands in the surrogate endpoints setting such as Example 1 could be encoded using the surrogate adjustment.

Combination of Surrogates and Mediators We now exemplify the capability of the composition of weighting operators in encoding more complicated causal effects. To do so, we explore the scenario combining Example 1 and 2 by permitting existence of the surrogates and mediators. Consider Fig. 1b in Appendix, where X represents a surrogate endpoint (e.g., CD4 cell counts), Y a clinical endpoint (e.g., survival from HIV), and Z a mediator between (X, Y) (e.g., a progression of HIV). W_1 and W_2 is set identical to Example 1.

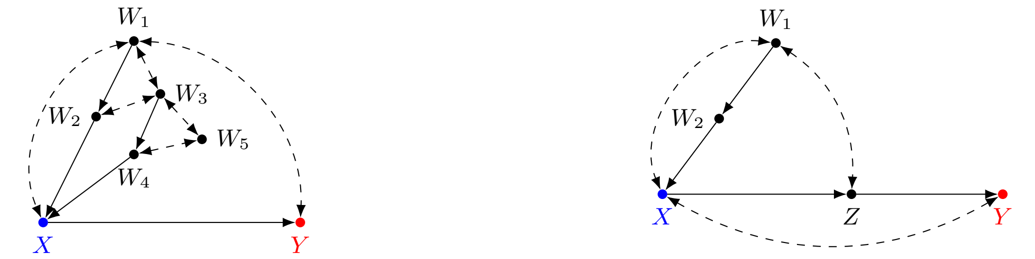
To identify the causal effect $P_x(y)$, we first note that Z satisfies the decomposability criteria (Def. 6). This leads to decompose the causal effect as $P_x(y) = \sum_z P_x(z)P_z(y)$. As alluded earlier, $\mathbb{E}_{P_x(z)}\left[h_z\left(Z\right)\right]$ can be identifiable by surrogate adjustment and given by Eq. (10).

To identify $P_z(y)$, defined by the effect of the mediator Z on the clinical endpoint Y, we first witness that Z is a mediator of (X, Y) where the spurious effect of Z on Y are blocked by X on the surrogate experiments on W_2; i.e., X satisfies BD criterion relative to (Z, Y) on the surrogate experimental distribution $P_{w_2}(y, x, z, w_1)$ and corresponding graph $G_{\overline{W_2}}$. This leads that $P_z(y)$ coincides with the weighted distribution $P^{\mathcal{W}_2}_{w_2}(y|z)$ where $P^{\mathcal{W}_2}_{w_2}(v)$ is a distribution weighting $P_{w_2}(y, x, z, w_1)$ by $\mathcal{W}_2 = P_{w_2}(z)/P_{w_2}(z|x)$. Notice that the surrogate experimental distribution $P_{w_2}(y, x, z, w_1)$ again coincides with the weighted distribution $P^{\mathcal{W}_1}(y, x, z, w_1|w_2)$ weighting $P(\mathbf{v})$ by $\mathcal{W}_1 = P(w_2)/P(w_2|w_1)$. This yields that the conditional distribution of Y given $\{W_2, Z\}$ weighted by $\mathcal{W} = \mathcal{W}_1 \times \mathcal{W}_2$ coincides with the $P_z(y)$:

$$\mathbb{E}_{P_z(y)}\left[Y\right] = \mathcal{B}\left[Y \mid \{w_2, z\}\,; \mathcal{W} = \mathcal{W}_1 \times \mathcal{W}_2\right]. \tag{2}$$

This leads to decompose the causal effect as $\mathbb{E}_{P_x(y)}\left[Y\right] = \left(\mathcal{B}_1 \circ \mathcal{B}_2\right)(x)$ where $\mathcal{B}_1\left(x\right)$ and $\mathcal{B}_2\left(z\right)$ are given by the surrogate adjustment in Eq. (10) and $\mathcal{B}_2\left(z\right) = \mathcal{B}\left[Y \mid \{w_2, z\}\,; \mathcal{W} = \mathcal{W}_1 \times \mathcal{W}_2\right]$, respectively

A caveat is that the values of Eq. (10,2) are independent to the value of W_2 by the *testable implication* of the causal diagram (Tian and Pearl 2002). This permits to write a causal effect as $\mathbb{E}_{P_x(y)}\left[Y\right] = \left(\mathcal{B}_1 \circ \mathcal{B}_2\right)(x)$ where $\mathbb{E}_{P_x(z)}\left[h_z\left(Z\right)\right] = \mathcal{B}_1\left(x\right)$ as in Eq. (10) and $\mathbb{E}_{P_z(y)}\left[Y\right] = \mathcal{B}_2\left(z\right)$ as in Eq. (2) in Appendix.



(a) Surrogate-endpoints with mSBD adjustment

(b) Combination of surrogates and mediators

Figure 1: An example graph **(a)** the surrogate criterion where $((W_2, W_4), (W_1, W_3))$ is surrogate admissible relative to (X, Y) ; and **(b)** Z satisfies a decomposability criteria and $P_x(z)$ is identifiable through the surrogate adjustment.

B Discussion of Sample Complexity of the CWO

Here we provide some intuitions why the CWO approach outperforms the naive one in the high-dimensional settings. Reliably computing conditional probabilities per stratum through the naive procedure is challenging since too few samples are usually available per stratum in high dimensional settings. In particular, Hoeffding's inequality says that for a binary random variable X, the minimum number of samples to acquire $(1-\alpha)$-confidence interval $\left(\frac{1}{N}\sum_{i=1}^{N} X_{(i)}\right) \pm t$ is $\frac{\log(2/\alpha)}{2t^2}$. For example, one might need approximately 185 samples drawn from $P(x|\mathbf{z})$ to obtain a reliable interval ranging ± 0.1 with 95% confidence given individual stratum $\mathbf{z} = (z_1, \cdots, z_d)$. This observation suggests that $O\left(2^d\right)$ samples (e.g., 10^6 samples if $d = 10$) are needed to reliably estimate $P\left(x|\mathbf{z}\right)$. Note that given the sample size N, the computational complexity of estimating conditional probabilities is given as $O\left(N2^d\right)$.

In contrast to the naive procedure, the proposed CWO estimator achieves more amenable complexities through the modeling-based approach and weighting techniques. This requires a correctly specified model class for $P\left(x|\mathbf{z}\right)$ (i.e., the modelled $\widehat{P}(x|\mathbf{z})$ is a consistent estimator of $P\left(x|\mathbf{z}\right)$); then much fewer samples are needed to reliable estimate the corresponding parameters. For example, a line of research (Peduzzi *et al.* 1996; Vittinghoff and McCulloch 2007; Austin and Steyerberg 2017) states that the reliable logistic parameters are derivable even when the number of samples per each variable $z_i \in \mathbf{z}$ (called the effects per variable, EPV) is 20-30 (i.e., 20×10 samples when $d = 10$). Since a wide class of regression methods take time polynomial in the sample size N, computationally, this method will be more efficient than its corresponding naive estimator. Obviously, selecting a sensible parameterization still represents a non-trivial challenge.

C Comparison with the Parametric Plug-in Estimator

In this section, we test the proposed method against the parametric plug-in estimator.

C.1 Simulation Setup

We estimate $\mu(\mathbf{x})$ by computing the mean of Y in $D_{int}(\mathbf{v})$, which is treated as the ground truth. We compare the proposed estimators with a parametric plug-in procedure, as discussed next:

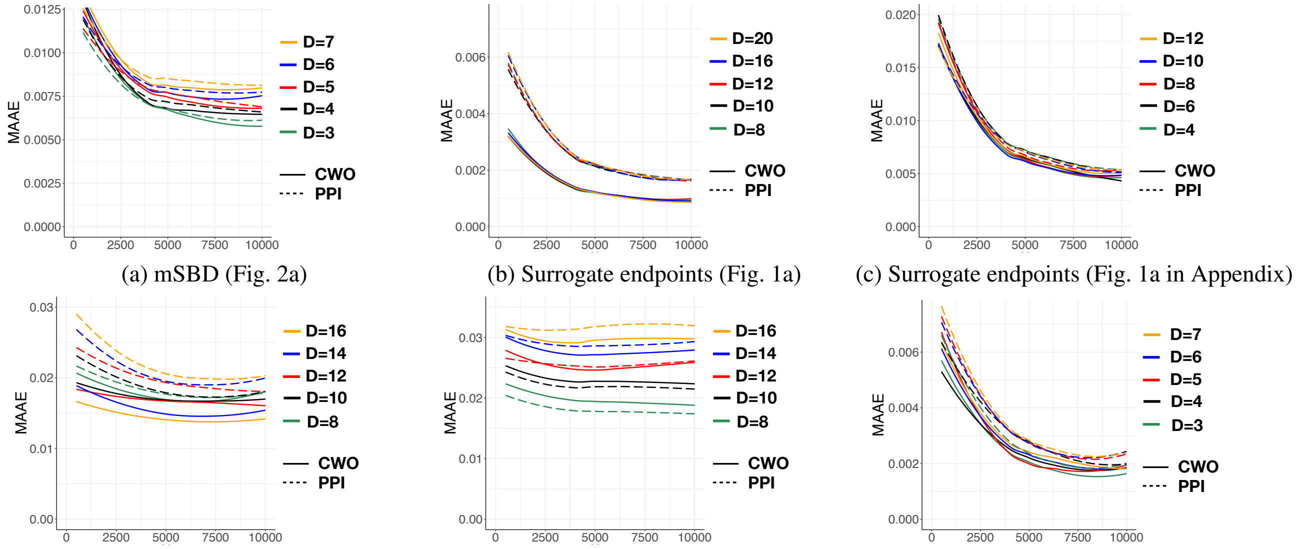
Parametric plug-in (shortly, PPI) Estimators As an example, assume we want to evaluate the causal estimand of the front-door adjustment. We estimate each conditional probabilities $\mathbb{E}\left[Y|x,z\right]$, $P(z|x)$ and $P(x')$ by imposing a parametric assumption (e.g., a logistic or linear regression). The expected causal effect is then computed by explicitly computing the estimand given estimated conditional probabilities; i.e., $\mathbb{E}_{P_x(y)}\left[Y\right] = \sum_z P(z|x)\sum_{x'}\mathbb{E}\left[Y|x',z\right]P(x')$. Notice that the PPI procedure is the only available parametric counterpart for comparison since no estimators have been developed for non-BDs estimands.

C.2 Simulation Results

We test the proposed CWO against the PPI procedure in aforementioned scenarios. Detailed descriptions of the corresponding SCMs are provided in Appendix E.

mSBD: (Fig. 2a) We test on estimating $\mathbb{E}_{P_{x_1,x_2}(y_2)}\left[Y_2\right]$, for which CWO and PPI procedures use results from Prop. 1 and Thm. 2 respectively. We set X_1, X_2, Y_1 to be binary, Y_2 continuous within $[0,1]$, and $Z_i = (Z_{i1}, \ldots, Z_{iD})$ for $i = 1, 2$, where all Z_{ij} are binary. Figure 2a presents the MAAE plots for $D = 3, 4, 5, 6, 7$. We note that CWO provides more reliable estimates compared to the PPI estimates. □

Surrogate endpoints (Fig. 1a) We test on estimating $\mathbb{E}_{P_x(y)}\left[Y\right]$ (where $P_x(y)$ given in Eq. (9)), which illustrates Example 1. We set W_2, X are binary and $Y \in [0,1]$ and $W_1 = (W_{11}, \cdots W_{1D})$ a set of D-dimensional binary variables. The MAAE plots for $D \in \{8, 10, 12, 16, 20\}$ are given in Fig. 2b. The CWO provides outperforms the PPI estimates. □



(a) mSBD (Fig. 2a)

(b) Surrogate endpoints (Fig. 1a)

(c) Surrogate endpoints (Fig. 1a in Appendix)

(d) Front-door (Fig. 2b)

(e) Causal mediators (Fig. 1b)

(f) Surrogates & mediators (Fig. 1b in Appendix)

Figure 2: MAAE plots comparing the proposed method with the parametric plug-in (PPI) procedure for **(a)** mSBD (Fig. 2a), **(b)** Surrogate endpoints (Fig. 1a), **(c)** Surrogate endpoints (Fig. 1a), **(d)** Causal mediators (Front-door, Fig. 2b), **(e)** Causal mediators with confounders (Fig. 1b), and **(f)** Combinations of surrogates and mediators (Fig. 1b). Plots are rendered in high resolution and can be zoomed in. Best viewed in color.

Surrogates endpoints with mSBD adjustment (Fig. 1a in Appendix) We test on estimating $\mathbb{E}_{P_x(y)}[Y]$ (where $P_x(y)$ is given in Eq. (1)). W_1, W_3 are set to be D-dimensional binary variables and $Y \in [0,1]$. Other variables are set to be binary. The MAAE plots for $D \in \{4,6,8,10,12\}$ are given in Fig. 2c. We observe that the result of the CWO is compatible with the PPI procedure. □

Front-door (Fig. 2b) We test on the front-door graph. To explore beyond Example 2, we set $\mathbf{Z}$ a D-dimensional binary variable. X is set to be binary, and Y continuous within $[0,1]$. The MAAE plots for $D \in \{8,10,12,14,16\}$ are provided in Fig. 2d. We observe that the result of the CWO estimator is more reliable compared to that of the PPI procedure. □

Causal mediators (Fig. 1b) We ran a test on estimating $\mathbb{E}_{P_x(y)}[Y]$ where $P_x(y)$ is given in Eq. 18. We set X, Z_1, Z_2, Z_3 to be binary, Y within $[0,1]$, $Z_4 = (Z_{4,1}, \ldots, Z_{4,D})$ a set of D-dimensional binary variable. The MAAE plots for $D \in \{8,10,12,14,18\}$ are provided in Fig. 2e. The plots imply that the CWO estimator provides compatible estimator to the PPI. □

Combinations of surrogates and mediators (Fig. 1b in Appendix) We simulated on estimating $\mathbb{E}_{P_x(y)}[Y]$, where the estimand is provided in Appendix A. We set W_2 and X as binary, W_1 and Z as D-dimensional binaries, and $Y \in [0,1]$. The MAAE plots for $D \in \{3,4,5,6,7\}$ are provided in Fig. 2f. The plots imply that the results of the CWO estimator are compatible with those of the PPI procedure. □

Overall, the CWO performs at least on par with the PPI, the only available counterparts for comparison on non-BDs estimands. The significance of the proposed estimator stems from that it provides a natural extension of well-known weighting based estimators (MSMs) to more general causal estimands. This extension provides a clue to generalize existing knowledge on BDs, such as the efficiency theory (Robins 1997), to non-BDs estimands. See Appendix C.3 for a detailed discussion on the competitive advantages of the CWO over the PPI.

C.3 Competitive Advantages of the CWO over the PPI estimator

This paper provides an approach of estimating a series of non-BDs estimands through a serial application of weighting based estimators (i.e., marginal structural models, MSMs (Robins *et al.* 2000)). A natural question that arises here is about competitive advantages of the proposed estimator compared to the parametric plug-in estimator procedure (PPI).

The fact that complicated causal estimands could be encoded through weighting operators is significant in many aspects. First, the approach of using MSMs to estimate non-BDs causal estimands is in principle applicable to binary, categorical, and continuous variables (VanderWeele 2009). For example, consider the case where the estimand of interest is given by the front-door adjustment . The proposed approach provides identical/unified procedures in estimating the estimand regardless of the mediators $\mathbf{Z}$ being discrete or continuous; the estimands are estimated through a chain of applications of MSMs. Whereas, the PPI procedure is not equipped with such computational convenience since one should employ heuristic integration technique to compute the marginalizing operator if $\mathbf{Z}$ is high-dimensional continuous variables.

Another advantage of the proposed method is that scientists are permitted to employ flexible and state-of-the-art machine learning methods developed for MSMs in estimating the causal estimands. For example, Gaussian Processes, well-known nonparametric regression techniques for arbitrary functions, are recently proposed as good candidate function classes in estimating the weighting operator (Wen *et al.* 2018). As another example, the modern deep learning algorithm has been developed for estimating the MSMs (Lim *et al.* 2018). Since our work provides a unified view of encoding complicated causal estimands using MSMs, the machine-learning-based MSMs estimators could be naturally extended toward estimators of more complicated causal instances.

Also, the proposed method bridges a well-known causal instance (BD/SBD) to the non-BD instances through novel machinery (i.e., the composition of weighting operators). The importance stems from that it paves the way toward a unified framework to construct estimators general identifiable causal instances. Notice that constructing estimators for every identifiable causal instance would be infeasible without the knowledge of the relationship between BD and non-BD. For example, we acknowledge a recent work (Fulcher *et al.* 2019) which constructs an efficient influence function and corresponding doubly-robust estimator for the causal instances satisfying front-door criteria. However, its relationship with the well-known instance (BD/SBD) is hardly appreciated in work, leading that generalization of the work challenging. Given that one can derive the influence function for the front-door adjustment by using simply summing two BD influence functions (due to Decomposability criteria in Def. 6), our work opens new research directions that the proposed machinery helps to transfer existing knowledge on BD to the non-BD settings.

Lastly, our method provides a semiparametric estimator for the causal estimand beyond BDs, which is agnostic in choosing the outcome model when the model is saturated (Robins 2000; Hernan and Robins 2019). Not only that, the proposed machinery (i.e., composition) in our work permits to extend the knowledge about the semiparametric efficiency bound on BDs toward non-BDs estimands.

D Proofs

Proof of Prop. 1 We first prove for BD adjustment. Note $P(\mathbf{x},\mathbf{y},\mathbf{z}) = P(\mathbf{y}|\mathbf{x},\mathbf{z})\,P(\mathbf{x}|\mathbf{z})\,P(\mathbf{z})$ by the chain rule of the probability. Given $\mathcal{W} \equiv \frac{P(\mathbf{x})}{P(\mathbf{x}|\mathbf{z})}$. Since $\sum_{\mathbf{x},\mathbf{y},\mathbf{z}} \mathcal{W}P(\mathbf{x},\mathbf{y},\mathbf{z}) = 1$, $P^{\mathcal{W}}(\mathbf{x},\mathbf{y},\mathbf{z}) = \mathcal{W}P(\mathbf{x},\mathbf{y},\mathbf{z}) = P(\mathbf{y}|\mathbf{x},\mathbf{z})\,P(\mathbf{x})\,P(\mathbf{z})$. Also, note that $P^{\mathcal{W}}(\mathbf{x}) = P(\mathbf{x})$, since $P^{\mathcal{W}}(\mathbf{x}) = \sum_{\mathbf{y},\mathbf{z}} P^{\mathcal{W}}(\mathbf{x},\mathbf{y},\mathbf{z}) = P(\mathbf{x})$. Therefore,

$$\begin{aligned}
P_{\mathbf{x}}(\mathbf{y}) &= \sum_{\mathbf{z}} P(\mathbf{y}|\mathbf{x},\mathbf{z})\,P(\mathbf{z}) \\
&= \sum_{\mathbf{z}} P(\mathbf{y}|\mathbf{x},\mathbf{z})\,\frac{P(\mathbf{x})}{P(\mathbf{x})}P(\mathbf{z}) \\
&= \frac{1}{P(\mathbf{x})}\sum_{\mathbf{z}} P(\mathbf{y}|\mathbf{x},\mathbf{z})\,P(\mathbf{x})\,P(\mathbf{z}) \\
&= \frac{1}{P(\mathbf{x})}\sum_{\mathbf{z}} P^{\mathcal{W}}(\mathbf{x},\mathbf{y},\mathbf{z}) \\
&= \frac{1}{P(\mathbf{x})}P^{\mathcal{W}}(\mathbf{x},\mathbf{y}) \\
&= \frac{1}{P^{\mathcal{W}}(\mathbf{x})}P^{\mathcal{W}}(\mathbf{x},\mathbf{y}) \\
&= P^{\mathcal{W}}(\mathbf{y}|\mathbf{x})\,.
\end{aligned}$$

We omit the proof for the SBD adjustment since it is implied by the proof of Thm. 2. By the definition of the expectation operator $\mathbb{E}$, it is obvious that $\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[\mathbf{Y}] = \mathbb{E}_{P^{\mathcal{W}}(\mathbf{y}|\mathbf{x})}[\mathbf{Y}|\mathbf{X}=\mathbf{x}]$. □

Proof of Thm. 1 Throughout the proof, we are marking $do(x)$ as $\widehat{x}$. Also, note $\mathbf{X}^{(i)} = \emptyset$ if $i \leq 0$. Suppose $\mathbf{Z}$ satisfies the mSBD conditions. Then the following holds:

$$\begin{aligned}
& P\left(\mathbf{y}|\widehat{\mathbf{x}}\right) \\
& = \sum_{\mathbf{z}_1} P\left(\mathbf{y}|\widehat{\mathbf{x}}, \mathbf{z}_1\right) P\left(\mathbf{z}_1\right) && \text{since } \mathbf{Z}_1 \subseteq ND\left(\mathbf{X}^{\geq 1}\right) \\
& = \sum_{\mathbf{z}_1} P\left(\mathbf{y}|\widehat{\mathbf{x}^{\geq 2}}, x_1, \mathbf{z}_1\right) P\left(\mathbf{z}_1\right) && \text{since } \left(\mathbf{Y} \perp\!\!\!\perp X_1|\mathbf{Z}_1\right)_{\underline{X_1}\overline{\mathbf{X}^{\geq 2}}} \qquad (3) \\
& = \sum_{\mathbf{z}_1} P\left(\mathbf{y}^{\geq 2}|\mathbf{y}_1, \mathbf{z}_1, x_1, \widehat{\mathbf{x}^{\geq 2}}\right) P\left(\mathbf{y}_1|\mathbf{z}_1, x_1\right) P\left(\mathbf{z}_1\right) && \text{where } \mathbf{Y}^{\geq 2} = \mathbf{Y} \cap De\left(\mathbf{X}^{\geq 2}\right) \qquad (4)
\end{aligned}$$

Note Eq. (3) holds by Rule 2 of *do*-calculus. Please refer (Pearl 2000, Eq. (4.5)). Suppose the following holds for $i \geq 2$:

$$\begin{aligned}
& P\left(\mathbf{y}|\widehat{\mathbf{x}}\right) \\
& = \sum_{\mathbf{z}^{(i-1)}} P\left(\mathbf{y}^{\geq i}|\mathbf{y}^{(i-1)}, \mathbf{z}^{(i-1)}, \mathbf{x}^{(i-1)}, \widehat{\mathbf{x}^{\geq i}}\right) \prod_{k=1}^{i-1} P\left(\mathbf{y}_k|\mathbf{z}^{(k)}, \mathbf{x}^{(k)}, \mathbf{y}^{(k-1)}\right) \prod_{j=1}^{i-1} P\left(\mathbf{z}_j|\mathbf{y}^{(j-1)}, \mathbf{x}^{(j-1)}, \mathbf{z}^{(j-1)}\right). \qquad (5)
\end{aligned}$$

Note that we already checked it Eq. (5) holds for $i = 2$. From the inductive hypothesis, the following holds:

$$\begin{aligned}
& P\left(\mathbf{y}|\widehat{\mathbf{x}}\right) \\
& = \sum_{\mathbf{z}^{(i-1)}} P\left(\mathbf{y}^{\geq i}|\mathbf{y}^{(i-1)}, \mathbf{z}^{(i-1)}, \mathbf{x}^{(i-1)}, \widehat{\mathbf{x}^{\geq i}}\right) \prod_{k=1}^{i-1} P\left(\mathbf{y}_k|\mathbf{z}^{(k)}, \mathbf{x}^{(k)}, \mathbf{y}^{(k-1)}\right) \prod_{j=1}^{i-1} P\left(\mathbf{z}_j|\mathbf{y}^{(j-1)}, \mathbf{x}^{(j-1)}, \mathbf{z}^{(j-1)}\right) \\
& = \sum_{\mathbf{z}^{(i-1)}, \mathbf{z}_i} P\left(\mathbf{y}^{\geq i}|\mathbf{y}^{(i-1)}, \mathbf{z}^{(i)}, \mathbf{x}^{(i-1)}, \widehat{\mathbf{x}^{\geq i}}\right) \prod_{k=1}^{i-1} P\left(\mathbf{y}_k|\mathbf{z}^{(k)}, \mathbf{x}^{(k)}, \mathbf{y}^{(k-1)}\right) \prod_{j=1}^{i-1} P\left(\mathbf{z}_j|\mathbf{y}^{(j-1)}, \mathbf{x}^{(j-1)}, \mathbf{z}^{(j-1)}\right) \\
& \times P\left(\mathbf{z}_i|\mathbf{y}^{(i-1)}, \mathbf{z}^{(i-1)}, \mathbf{x}^{(i-1)}\right) \text{ by } \mathbf{Z}_i \subseteq ND\left(\mathbf{x}^{\geq i}\right) \\
& = \sum_{\mathbf{z}^{(i)}} P\left(\mathbf{y}^{\geq i}|\mathbf{y}^{(i-1)}, \mathbf{z}^{(i)}, \mathbf{x}^{(i)}, \widehat{\mathbf{x}^{\geq i+1}}\right) \prod_{k=1}^{i-1} P\left(\mathbf{y}_k|\mathbf{z}^{(k)}, \mathbf{x}^{(k)}, \mathbf{y}^{(k-1)}\right) \prod_{j=1}^{i} P\left(\mathbf{z}_j|\mathbf{y}^{(j-1)}, \mathbf{x}^{(j-1)}, \mathbf{z}^{(j-1)}\right) \qquad (6) \\
& = \sum_{\mathbf{z}^{(i)}} P\left(\mathbf{y}^{\geq i+1}|\mathbf{y}^{(i)}, \mathbf{z}^{(i)}, \mathbf{x}^{(i)}, \widehat{\mathbf{x}^{\geq i+1}}\right) \prod_{k=1}^{i} P\left(\mathbf{y}_k|\mathbf{z}^{(k)}, \mathbf{x}^{(k)}, \mathbf{y}^{(k-1)}\right) \prod_{j=1}^{i} P\left(\mathbf{z}_j|\mathbf{y}^{(j-1)}, \mathbf{x}^{(j-1)}, \mathbf{z}^{(j-1)}\right), \qquad (7)
\end{aligned}$$

where Eq. (6) holds by the condition $\left(\mathbf{Y}^{\geq i} \perp\!\!\!\perp X_i|\mathbf{Y}^{(i-1)}, \mathbf{Z}^{(i)}, \mathbf{X}^{(i)}\right)_{\underline{X_i}\overline{\mathbf{X}^{\geq i+1}}}$, and Eq. (7) holds by $\mathbf{Y}_i \subseteq ND\left(\mathbf{X}^{\geq i+1}\right)$. This shows that the inductive hypothesis holds for all i.

Note that we can consider when $\mathbf{Y}_0 = \mathbf{Y} \setminus De\left(\mathbf{X}\right)$ without loss of generality since the proof holds by starting with $P\left(\mathbf{y}|\widehat{\mathbf{x}}\right) = P\left(\mathbf{y}_0\right) P\left(\mathbf{y}_1|\mathbf{y}_0, \widehat{\mathbf{x}}\right)$. This completes the proof. □

Proof of Thm. 2 We define some useful quantities as follow:

$$\begin{aligned}
q\left(x_k|\mathbf{z}^{(k)}, \mathbf{y}^{(k-1)}\right) &\equiv P\left(x_k|\mathbf{x}^{(k-1)}, \mathbf{z}^{(k)}, \mathbf{y}^{(k-1)}\right) \\
q\left(y_k|\mathbf{x}^{(k)}, \mathbf{z}^{(k)}\right) &\equiv P\left(y_k|\mathbf{y}^{(k-1)}, \mathbf{x}^{(k)}, \mathbf{z}^{(k)}\right) \\
q\left(z_k|\mathbf{x}^{(k-1)}, \mathbf{y}^{(k-1)}\right) &\equiv P\left(z_k|\mathbf{z}^{(k-1)}, \mathbf{x}^{(k-1)}, \mathbf{y}^{(k-1)}\right)
\end{aligned}$$

and

$$\begin{aligned}
q(\mathbf{x}|\mathbf{y},\mathbf{z}) &\equiv \prod_{k=1}^{n} q\left(x_k|\mathbf{z}^{(k)},\mathbf{y}^{(k-1)}\right) \\
&= \prod_{k=1}^{n} P\left(x_k|\mathbf{x}^{(k-1)},\mathbf{z}^{(k)},\mathbf{y}^{(k-1)}\right) \\
q\left(\mathbf{y}|\mathbf{x},\mathbf{z}\right) &\equiv \prod_{k=1}^{n} q\left(y_k|\mathbf{x}^{(k)},\mathbf{z}^{(k)}\right) \\
&= \prod_{k=1}^{n} P\left(y_k|\mathbf{y}^{(k-1)},\mathbf{x}^{(k)},\mathbf{z}^{(k)}\right) \\
q\left(\mathbf{z}|\mathbf{x},\mathbf{y}\right) &\equiv \prod_{k=1}^{n} q\left(z_k|\mathbf{x}^{(k-1)},\mathbf{y}^{(k-1)}\right) \\
&= \prod_{k=1}^{n} P\left(z_k|\mathbf{z}^{(k-1)},\mathbf{x}^{(k)},\mathbf{y}^{(k)}\right).
\end{aligned}$$

First, it is immediate to witness that $P(\mathbf{x},\mathbf{y},\mathbf{z}) = q(\mathbf{x}|\mathbf{y},\mathbf{z})q(\mathbf{y}|\mathbf{x},\mathbf{z})q(\mathbf{z}|\mathbf{y},\mathbf{z})$, since $P\left(\mathbf{x},\mathbf{y},\mathbf{z}\right) = P\left(z_1\right)P\left(x_1|z_1\right)P\left(y_1|x_1,z_1\right)P\left(z_2|x_1,z_1,y_1\right)\cdots P\left(y_n|\mathbf{x}^{(n)},\mathbf{y}^{(n-1)},\mathbf{z}^{(n)}\right)$. Also, we can rewrite the weight as $\mathcal{W} = \mathcal{W}(\mathbf{x},\mathbf{y},\mathbf{z}) = \frac{P(\mathbf{x})}{q(\mathbf{x}|\mathbf{y},\mathbf{z})}$, which leads that $P^{\mathcal{W}}\left(\mathbf{x},\mathbf{y},\mathbf{z}\right) = \mathcal{W}P\left(\mathbf{x},\mathbf{y},\mathbf{z}\right) = q\left(\mathbf{y}|\mathbf{x},\mathbf{z}\right)q\left(\mathbf{z}|\mathbf{x},\mathbf{y}\right)P\left(\mathbf{x}\right)$. Also, $P^{\mathcal{W}}\left(\mathbf{x}\right) = \sum_{\mathbf{y},\mathbf{z}} P^{\mathcal{W}}\left(\mathbf{x},\mathbf{y},\mathbf{z}\right) = \sum_{\mathbf{y},\mathbf{z}} q\left(\mathbf{y}|\mathbf{x},\mathbf{z}\right)q\left(\mathbf{z}|\mathbf{x},\mathbf{y}\right)P\left(\mathbf{x}\right) = P\left(\mathbf{x}\right)$. Finally, $P_{\mathbf{x}}\left(\mathbf{y}\right) = \sum_{\mathbf{z}} q\left(\mathbf{y}|\mathbf{x},\mathbf{z}\right)q(\mathbf{z}|\mathbf{y},\mathbf{z})$.

Then the causal effect can be rewritten as follow:

$$\begin{aligned}
P_{\mathbf{x}}\left(\mathbf{y}\right) &= \sum_{\mathbf{z}} q(\mathbf{y}|\mathbf{x},\mathbf{z})q(\mathbf{z}|\mathbf{y},\mathbf{x}) \\
&= \sum_{\mathbf{z}} q(\mathbf{y}|\mathbf{x},\mathbf{z})q(\mathbf{z}|\mathbf{y},\mathbf{x})\frac{q(\mathbf{x}|\mathbf{y},\mathbf{z})}{q(\mathbf{x}|\mathbf{y},\mathbf{z})} \\
&= \sum_{\mathbf{z}} P\left(\mathbf{x},\mathbf{y},\mathbf{z}\right)\frac{1}{q\left(\mathbf{x}|\mathbf{y},\mathbf{z}\right)} \\
&= \sum_{\mathbf{z}} P\left(\mathbf{x},\mathbf{y},\mathbf{z}\right)\frac{1}{q\left(\mathbf{x}|\mathbf{y},\mathbf{z}\right)}\frac{P(\mathbf{x})}{P\left(\mathbf{x}\right)} \\
&= \sum_{\mathbf{z}} \frac{1}{P(\mathbf{x})}P(\mathbf{x},\mathbf{y},\mathbf{z})\frac{P(\mathbf{x})}{q(\mathbf{x}|\mathbf{y},\mathbf{z})} \\
&= \frac{1}{P(\mathbf{x})}\sum_{\mathbf{z}} \mathcal{W}P(\mathbf{x},\mathbf{y},\mathbf{z}) \\
&= \frac{1}{P(\mathbf{x})}\sum_{\mathbf{z}} P^{\mathcal{W}}\left(\mathbf{x},\mathbf{y},\mathbf{z}\right) \\
&= \frac{1}{P(\mathbf{x})}P^{\mathcal{W}}\left(\mathbf{x},\mathbf{y}\right) \\
&= \frac{1}{P^{\mathcal{W}}\left(\mathbf{x}\right)}P^{\mathcal{W}}\left(\mathbf{x},\mathbf{y}\right) = P^{\mathcal{W}}\left(\mathbf{y}|\mathbf{x}\right),
\end{aligned}$$

which completes the proof that $P_{\mathbf{x}}\left(\mathbf{y}\right) = P^{\mathcal{W}}\left(\mathbf{y}|\mathbf{x}\right)$ given $\mathcal{W}$. By definition of the expectation operator, this completes the proof. □

Proof of Thm. 3 Suppose $(\mathbf{R},\mathbf{Z})$ is surrogate admissible relative to $(\mathbf{X},\mathbf{Y})$. Then the causal effect $P_{\mathbf{x}}\left(\mathbf{y}\right)$ is given as follow:

$$\begin{aligned}
P_{\mathbf{x}}\left(\mathbf{y}\right) &= P_{\mathbf{x},\mathbf{r}}\left(\mathbf{y}\right) && \text{By Rule 3 of } do\text{-calculus from } (\mathbf{Y} \perp\!\!\!\perp \mathbf{R}|\mathbf{X})_{G_{\overline{\mathbf{X}\mathbf{R}}}} \\
&= P_{\mathbf{r}}\left(\mathbf{y}|\mathbf{x}\right) && \text{By Rule 2 of } do\text{-calculus from } (\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{R})_{G_{\underline{\mathbf{X}}\overline{\mathbf{R}}}} \\
&= \frac{P_{\mathbf{r}}\left(\mathbf{y},\mathbf{x}\right)}{P_{\mathbf{r}}\left(\mathbf{x}\right)}. \qquad (8)
\end{aligned}$$

Since $\mathbf{Z}$ is mSBD admissible relative to $(\mathbf{R}, (\mathbf{X}, \mathbf{Y}))$, the weights is given by $\mathcal{W} \equiv \mathcal{W}_{\text{mSBD}}\left(\mathbf{r}, \mathbf{x} \cup \mathbf{y}, \mathbf{z}\right)$ by Thm. 2, and $P^{\mathcal{W}}\left(\mathbf{x}, \mathbf{y}|\mathbf{r}\right) = P_{\mathbf{r}}\left(\mathbf{x}, \mathbf{y}\right)$. Also, $P^{\mathcal{W}}\left(\mathbf{x}|\mathbf{r}\right) = P_{\mathbf{r}}\left(\mathbf{x}\right)$ by $\mathbf{Z}$ surrogate-admissibility. Therefore, Eq. (8) can be rewritten as follow:

$$P_{\mathbf{x}}\left(\mathbf{y}\right) = \frac{P_{\mathbf{r}}\left(\mathbf{y}, \mathbf{x}\right)}{P_{\mathbf{r}}\left(\mathbf{x}\right)} = \frac{P^{\mathcal{W}}\left(\mathbf{x}, \mathbf{y}|\mathbf{r}\right)}{P^{\mathcal{W}}\left(\mathbf{x}|\mathbf{r}\right)} = P^{\mathcal{W}}\left(\mathbf{y}|\mathbf{x}, \mathbf{r}\right).$$

Then the causal effect $\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}\left[h\left(\mathbf{Y}\right)\right]$ is given as follow:

$$\begin{aligned}
\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}\left[h\left(\mathbf{Y}\right)\right] &= \sum_{\mathbf{y}} h\left(\mathbf{y}\right) P_{\mathbf{x}}\left(\mathbf{y}\right) \\
&= \sum_{\mathbf{y}} h\left(\mathbf{y}\right) P^{\mathcal{W}}\left(\mathbf{y}|\mathbf{x}, \mathbf{r}\right) \\
&= \mathcal{B}\left[h\left(\mathbf{Y}\right) \mid \mathbf{x} \cup \mathbf{r}; \mathcal{W} = \mathcal{W}_{\text{mSBD}}\left(\mathbf{r}, \mathbf{x} \cup \mathbf{y}, \mathbf{z}\right)\right],
\end{aligned}$$

which completes the proof. □

Proof of Prop. 2 By recalling the definition of the composition of weighting operators $\mathcal{B}_1$ and $\mathcal{B}_2$, $\left(\mathcal{B}_1 \circ \mathcal{B}_2\right)\left(\mathbf{x}\right)$, the composition is given by

$$\begin{aligned}
\left(\mathcal{B}_1 \circ \mathcal{B}_2\right)\left(\mathbf{x}\right) &= \mathcal{B}\left[\mathcal{B}_2\left(\mathbf{z}\right) \mid \mathbf{x}; \mathcal{W}_1\right] \\
&= \sum_{\mathbf{z}} \mathcal{B}_2\left(\mathbf{z}\right) P^{\mathcal{W}_1}\left(\mathbf{z}|\mathbf{x}\right) \\
&= \sum_{\mathbf{z}} \mathcal{B}_2\left(\mathbf{z}\right) P\left(\mathbf{z}|\mathbf{x}\right) \\
&= \sum_{\mathbf{z}} \mathcal{B}\left[\mathbf{Y} \mid \mathbf{z}; \mathcal{W}_2\right] P\left(\mathbf{z}|\mathbf{x}\right) \\
&= \sum_{\mathbf{z}} \left(\sum_{\mathbf{y}} \mathbf{y} P^{\mathcal{W}_2}\left(\mathbf{y}|\mathbf{z}\right)\right) P\left(\mathbf{z}|\mathbf{x}\right) \\
&= \sum_{\mathbf{z}} \left(\sum_{\mathbf{y}} \mathbf{y} \sum_{\mathbf{x}'} P\left(\mathbf{y}|\mathbf{x}', \mathbf{z}\right) P\left(\mathbf{x}'\right)\right) P\left(\mathbf{z}|\mathbf{x}\right) \qquad (9) \\
&= \sum_{\mathbf{z}} P\left(\mathbf{z}|\mathbf{x}\right) \sum_{\mathbf{x}'} \mathbb{E}\left[\mathbf{Y}|\mathbf{x}', \mathbf{z}\right] P\left(\mathbf{x}'\right). \qquad (10)
\end{aligned}$$

Since Eq. (10) coincides with the the front-door adjustment (Pearl 2000), this completes the proof.

To witness Eq. (9), consider the following weighted distribution:

$$\begin{aligned}
P^{\mathcal{W}_2}\left(\mathbf{x}, \mathbf{y}, \mathbf{z}\right) &= \frac{\mathcal{W}_2 P\left(\mathbf{x}, \mathbf{y}, \mathbf{z}\right)}{\sum_{\mathbf{x}, \mathbf{y}, \mathbf{z}} \mathcal{W}_2 P\left(\mathbf{x}, \mathbf{y}, \mathbf{z}\right)} \\
&= \mathcal{W}_2 P\left(\mathbf{x}, \mathbf{y}, \mathbf{z}\right) \\
&= \frac{P\left(\mathbf{z}\right)}{P\left(\mathbf{z}|\mathbf{x}\right)} P\left(\mathbf{y}|\mathbf{x}, \mathbf{z}\right) P\left(\mathbf{z}|\mathbf{x}\right) P\left(\mathbf{x}\right) \\
&= P\left(\mathbf{y}|\mathbf{x}, \mathbf{z}\right) P\left(\mathbf{x}\right) P\left(\mathbf{z}\right),
\end{aligned}$$

which leads the following:

$$P^{\mathcal{W}_2}\left(\mathbf{y}|\mathbf{z}\right) = \sum_{\mathbf{x}'} P\left(\mathbf{y}|\mathbf{x},'\mathbf{z}\right) P\left(\mathbf{x}'\right).$$

□

Proof of Thm. 4 Let $\mathbf{Z}$ hold the decomposability criterion $(\mathbf{x}, \mathbf{y})$. Then

$$\begin{aligned}
P_{\mathbf{x}}\left(\mathbf{y}\right) &= \sum_{\mathbf{z}} P_{\mathbf{x}}\left(\mathbf{y}|\mathbf{z}\right) P_{\mathbf{x}}\left(\mathbf{z}\right) && \text{Marginalizing over } \mathbf{Z} \\
&= \sum_{\mathbf{z}} P_{\mathbf{x}, \mathbf{z}}\left(\mathbf{y}\right) P_{\mathbf{x}}\left(\mathbf{z}\right) && \text{By } \left(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}\right)_{G_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}}\text{, Rule 2 of } do\text{-calculus} \\
&= \sum_{\mathbf{z}} P_{\mathbf{z}}\left(\mathbf{y}\right) P_{\mathbf{x}}\left(\mathbf{z}\right) && \text{By } \left(\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{Z}\right)_{G_{\overline{\mathbf{X}\mathbf{Z}}}}\text{, Rule 3 of } do\text{-calculus}\ .
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}\left[h\left(\mathbf{Y}\right)\right] &= \sum_{\mathbf{y}} h\left(\mathbf{y}\right) P_{\mathbf{x}}\left(\mathbf{y}\right) \\
&= \sum_{\mathbf{y}} h\left(\mathbf{y}\right) \sum_{\mathbf{z}} P_{\mathbf{z}}\left(\mathbf{y}\right) P_{\mathbf{x}}\left(\mathbf{z}\right) \\
&= \sum_{\mathbf{z}} P_{\mathbf{x}}\left(\mathbf{z}\right) \sum_{\mathbf{y}} h\left(\mathbf{y}\right) P_{\mathbf{z}}\left(\mathbf{y}\right) \\
&= \sum_{\mathbf{z}} \mathbb{E}_{P_{\mathbf{z}}(\mathbf{y})}\left[h\left(\mathbf{Y}\right)\right] \\
& \mathbb{E}_{P_{\mathbf{x}}(\mathbf{z})}\left[\mathbb{E}_{P_{\mathbf{z}}(\mathbf{y})}\left[h\left(\mathbf{Y}\right)\right]\right],
\end{aligned}$$

which completes the proof. □

Proof of Thm. 5 Suppose $(\mathbf{Z}, \mathbf{W}_1, \mathbf{W}_2)$ holds the SBD composition criterion. By $\mathbf{Z}$ satisfying the decomposability relative to $(\mathbf{x}, \mathbf{y})$,

$$\begin{aligned}
\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}\left[h\left(\mathbf{Y}\right)\right] &= \sum_{\mathbf{y}} h\left(\mathbf{y}\right) P_{\mathbf{x}}\left(\mathbf{y}\right) \\
&= \sum_{\mathbf{y}} h\left(\mathbf{y}\right) \sum_{\mathbf{z}} P_{\mathbf{z}}\left(\mathbf{y}\right) P_{\mathbf{x}}\left(\mathbf{z}\right) \\
&= \sum_{\mathbf{y},\mathbf{z}} h\left(\mathbf{y}\right) P_{\mathbf{z}}\left(\mathbf{y}\right) P_{\mathbf{x}}\left(\mathbf{z}\right) \\
&= \mathbb{E}_{P_{\mathbf{x}}(\mathbf{z})}\left[\sum_{\mathbf{y}} h\left(\mathbf{Y}\right) P_{\mathbf{z}}\left(\mathbf{y}\right)\right] \\
&= \mathbb{E}_{P_{\mathbf{x}}(\mathbf{z})}\left[\mathbb{E}_{P_{\mathbf{z}}(\mathbf{y})}\left[h\left(\mathbf{Y}\right)\right]\right].
\end{aligned}$$

Since $\mathbf{W}_1$ is mSBD admissible relative to $(\mathbf{X}, \mathbf{Z})$, the causal effect $\mathbb{E}_{P_{\mathbf{x}}(\mathbf{z})}\left[h_z\left(\mathbf{Z}\right)\right]$ is given by

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{z})}\left[h_z\left(\mathbf{Z}\right)\right] = \mathcal{B}_1\left(\mathbf{x}\right) \equiv \mathcal{B}\left[h_z\left(\mathbf{Z}\right) \mid \mathbf{x}; \mathcal{W}_1\right],$$

where $\mathcal{W}_1 \equiv \mathcal{W}_{\text{mSBD}}\left(\mathbf{x}, \mathbf{z}, \mathbf{w}_1\right)$.

In similar, $\mathbf{W}_2$ is mSBD admissible relative to $(\mathbf{Z}, \mathbf{Y})$, leading that the causal effect $\mathbb{E}_{P_{\mathbf{z}}(\mathbf{y})}\left[h_y\left(\mathbf{Y}\right)\right]$ is given by

$$\mathbb{E}_{P_{\mathbf{z}}(\mathbf{y})}\left[h_y\left(\mathbf{Y}\right)\right] = \mathcal{B}_2\left(\mathbf{z}\right) \equiv \mathcal{B}\left[h_y\left(\mathbf{Y}\right) \mid \mathbf{z}; \mathcal{W}_2\right],$$

where $\mathcal{W}_2 \equiv \mathcal{W}_{\text{mSBD}}\left(\mathbf{z}, \mathbf{y}, \mathbf{w}_2\right)$.

Since $\mathcal{B}_2\left(\mathbf{z}\right)$ is a function of $\mathbf{z}$, the composition of two operators $\mathcal{B}_1\left(\mathbf{x}\right)$ and $\mathcal{B}_2\left(\mathbf{z}\right)$, $\left(\mathcal{B}_1 \circ \mathcal{B}_2\right)\left(\mathbf{x}\right)$, is well-defined and given as

$$\begin{aligned}
\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}\left[h\left(\mathbf{Y}\right)\right] &= \mathbb{E}_{P_{\mathbf{x}}(\mathbf{z})}\left[\mathbb{E}_{P_{\mathbf{z}}(\mathbf{y})}\left[h\left(\mathbf{Y}\right)\right]\right] \\
&= \mathbb{E}_{P_{\mathbf{x}}(\mathbf{z})}\left[\mathcal{B}_2\left(\mathbf{z}\right)\right] \\
&= \mathcal{B}\left[\mathcal{B}_2\left(\mathbf{z}\right) \mid \mathbf{x}; \mathcal{W}_1\right] \\
&= \left(\mathcal{B}_1 \circ \mathcal{B}_2\right)\left(\mathbf{x}\right),
\end{aligned}$$

by the definition of weighting operators and the composition operators of weighting operators. This completes the proof. □

Proof of Thm. 6 Throughout the proof, we denote $\widehat{\mathcal{W}} \equiv \widehat{\mathcal{W}}\left(\mathbf{v}\right)$ and $\widehat{\mathcal{W}}_{(i)} \equiv \widehat{\mathcal{W}}\left(\mathbf{v}_{(i)}\right)$. Given estimated weights $\widehat{\mathcal{W}}_1$, let $\widehat{\mathcal{B}}_{1,N_1}\left(\mathbf{x}, h_z\left(\mathbf{z}\right)\right) \equiv \widehat{\mathcal{B}}\left[h_z\left(\mathbf{Z}\right) \mid \mathbf{x}; \mathcal{W}_1\right]$ denote the consistent estimate of $\mathcal{B}_1\left(\mathbf{x}, h_z\left(\mathbf{z}\right)\right) \equiv \mathcal{B}\left[h_z\left(\mathbf{Z}\right) \mid \mathbf{x}; \mathcal{W}_1\right]$ estimated from N_1 finite samples. In other words.

$$\widehat{\mathcal{B}}_{1,N_1}\left(\mathbf{x}, h_z\left(\mathbf{z}\right)\right) \equiv \arg\min_{g_1 \in \mathcal{R}} \sum_{i=1}^{N_1} \widehat{\mathcal{W}}_{1,(i)}\left(h_z\left(\mathbf{Z}_{(i)}\right) - \widehat{g}_1\left(\mathbf{X}_{(i)}\right)\right)^2. \tag{11}$$

In similar, given estimated $\widehat{\mathcal{W}_2}$, let $\widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_{\mathbf{y}}\left(\mathbf{y}\right)\right) \equiv \widehat{\mathcal{B}}\left[h_y\left(\mathbf{Y}\right) \mid \mathbf{z}; \mathcal{W}_2\right]$ denote the consistent estimate of $\mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right) \equiv \mathcal{B}\left[h_y\left(\mathbf{Y}\right) \mid \mathbf{z}; \mathcal{W}_2\right]$ estimated from N_2 finite samples. In other words.

$$\widehat{\mathcal{B}}_{1,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right) \equiv \arg\min_{g_2 \in \mathcal{R}} \sum_{i=1}^{N_2} \widehat{\mathcal{W}}_{2,(i)} \left(h_y\left(\mathbf{Y}_{(i)}\right) - \widehat{g}_2\left(\mathbf{Z}_{(i)}\right)\right)^2. \tag{12}$$

By the definition of the consistent estimator, whenever $N_1 \to \infty$, the function $\widehat{\mathcal{B}}_{1,N_1}\left(\mathbf{x}, h_z\left(\mathbf{z}\right)\right) \xrightarrow{P} \mathcal{B}_1\left(\mathbf{x}, h_z\left(\mathbf{z}\right)\right)$. Also, whenever $N_2 \to \infty$, $\widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right) \xrightarrow{P} \mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)$. That is, for any positive ϵ_1, ϵ_2 and δ_1, δ_2, for all $N_1 > N_1'$ and $N_2 > N_2'$ for some fixed N_1', N_2', we have

$$\begin{aligned} P\left(\left| \widehat{\mathcal{B}}_{1,N_1}\left(\mathbf{x}, h_z\left(\mathbf{z}\right)\right) - \mathcal{B}_1\left(\mathbf{x}, h_z\left(\mathbf{z}\right)\right) \right| > \epsilon_1\right) < \delta_1 \\ P\left(\left| \widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right) - \mathcal{B}_2\left(\mathbf{y}, h_y\left(\mathbf{y}\right)\right) \right| > \epsilon_2\right) < \delta_2, \end{aligned}$$

by the definition of *convergence in probability*.

Using such notations, the composition of $\widehat{\mathcal{B}}_{1,N_1}\left(\mathbf{x}, h_z\left(\mathbf{z}\right)\right)$ and $\widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)$ can be written as follow:

$$\left(\widehat{\mathcal{B}}_{1,N_1} \circ \widehat{\mathcal{B}}_{2,N_2}\right)\left(\mathbf{x}\right) = \widehat{\mathcal{B}}_{1,N_1}\left(\mathbf{x}, \widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)\right) \tag{13}$$

Now, we are going to show the consistency of the composition estimator by proving the following:

$$\left(\widehat{\mathcal{B}}_{1,N_1} \circ \widehat{\mathcal{B}}_{2,N_2}\right)\left(\mathbf{x}\right) \xrightarrow{P} \left(\mathcal{B}_1 \circ \widehat{\mathcal{B}}_2\right)\left(\mathbf{x}\right) = \mathcal{B}_1\left(\mathbf{x}, \mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)\right).$$

For convenience, we define the following quantities:

$$\begin{aligned} A &\equiv \widehat{\mathcal{B}}_{1,N_1}\left(\mathbf{x}, \widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)\right) \\ B &\equiv \widehat{\mathcal{B}}_{1,N_1}\left(\mathbf{x}, \mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)\right) \\ C &\equiv \mathcal{B}_1\left(\mathbf{x}, \widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)\right) \\ D &\equiv \mathcal{B}_1\left(\mathbf{x}, \mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)\right) \end{aligned}$$

Our goal is then to show $A \xrightarrow{P} D$. Equivalently, we need to show that for any positive ϵ and δ, for all $N_1 > N_1'$ and $N_2 > N_2'$ for some fixed N_1', N_2', $P\left(|A - D| > \epsilon\right) < \delta$ holds. To show, consider the following:

$$\begin{aligned} P\left(|A - D| > \epsilon\right) &= P\left(|A - B + B - D| > \epsilon\right) \\ &\leq P\left(|A - B| + |B - D| > \epsilon\right) \\ &\leq \left(\int_0^\epsilon P\left(|A - B| = t, |B - D| > \epsilon - t\right) dt\right) + P\left(|A - B| > \epsilon\right) \\ &\leq \left(\int_0^\epsilon P\left(|A - B| = t^*, |B - D| > \epsilon - t^*\right) dt\right) + P\left(|A - B| > \epsilon\right) \\ &= \epsilon P\left(|A - B| = t^*, |B - D| > \epsilon - t^*\right) + P\left(|A - B| > \epsilon\right) \\ &\leq \epsilon P\left(|B - D| > \epsilon - t^*\right) + P\left(|A - B| > \epsilon\right) \end{aligned} \tag{14}$$

where $t^* \equiv \arg\max_{t \in [0,\epsilon)} P\left(|A - B| = t, |B - D| > \epsilon - t\right)$.

Note that $\epsilon P\left(|B - D| > \epsilon - t^*\right)$ in Eq. (14) converges to 0 whenever $N_1 \to \infty$ since we assumed $B \xrightarrow{P} D$. Therefore, it is sufficient to show that for any ϵ and δ there exists N_1', N_2' such that $P\left(|A - B| > \epsilon\right) < \delta$ for all N_1, N_2.

We remind the definition of A and B. A precise definition of A is following:

$$\begin{aligned} A &\equiv \widehat{\mathcal{B}}_{1,N_1}\left(\mathbf{x}, \widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)\right) \\ &= \arg\min_{g_1 \in \mathcal{R}} \sum_{i=1}^{N_1} \widehat{\mathcal{W}}_{1,(i)} \left(\widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right) - \widehat{g}_1\left(\mathbf{X}_{(i)}\right)\right)^2 \\ &= \arg\min_{g_1 \in \mathcal{R}} \sum_{i=1}^{N_1} \widehat{\mathcal{W}}_{1,(i)} \left(\left(\widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)\right)^2 - 2\left(\widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)\right)\left(\widehat{g}_1\left(\mathbf{X}_{(i)}\right)\right) + \left(\widehat{g}_1\left(\mathbf{X}_{(i)}\right)\right)^2\right). \end{aligned}$$

In similar, the precise definition of B is following:

$$\begin{aligned}
B &\equiv \widehat{\mathcal{B}}_{1,N_1}\left(\mathbf{x}, \mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)\right) \\
&= \arg\min_{g_1 \in \mathcal{R}} \sum_{i=1}^{N_1} \widehat{\mathcal{W}}_{1,(i)} \left(\mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right) - \widehat{g_1}\left(\mathbf{X}_{(i)}\right)\right)^2 \\
&= \arg\min_{g_1 \in \mathcal{R}} \sum_{i=1}^{N_1} \widehat{\mathcal{W}}_{1,(i)} \left(\left(\mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)\right)^2 - 2\left(\mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)\right)\left(\widehat{g_1}\left(\mathbf{X}_{(i)}\right)\right) + \left(\widehat{g_1}\left(\mathbf{X}_{(i)}\right)\right)^2\right).
\end{aligned}$$

Note that A is dependent on the samples sizes N_1 and N_2 meanwhile B is only dependent on the sample size N_1 since B is already equipped with the $\mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)$.

Now, we define two quantities as follow:

$$\begin{aligned}
F_{N_2}\left(g_1\right) &\equiv \sum_{i=1}^{N_1} \widehat{\mathcal{W}}_{1,(i)} \left(\widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right) - \widehat{g_1}\left(\mathbf{X}_{(i)}\right)\right)^2 \\
F\left(g_1\right) &\equiv \sum_{i=1}^{N_1} \widehat{\mathcal{W}}_{1,(i)} \left(\mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right) - \widehat{g_1}\left(\mathbf{X}_{(i)}\right)\right)^2.
\end{aligned}$$

Using such definition, $A = \arg\min_{g_1 \in \mathcal{R}} F_{N_2}(g_1)$ and $B = \arg\min_{g_1 \in \mathcal{R}} F(g_1)$. Clearly, $F_{N_2}(g_1) \overset{P}{\to} F(g_1)$ since $\widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right) \overset{P}{\to} \mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)$. In particular, $F_{N_2}(g_1) \overset{P}{\to} F(g_1)$ is guaranteed by the *continuous mapping theorem* (Mann and Wald 1943); if $\widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right) \overset{P}{\to} \mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)$, then the continuous function of $\widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)$ also converges.

Given $F_{N_2}(g_1) \overset{P}{\to} F(g_1)$, we now need to show that the extremum of the quantity (i.e., $A = \arg\min_{g_1 \in \mathcal{R}} F_{N_2}(g_1)$ and $B = \arg\min_{g_1 \in \mathcal{R}} F(g_1)$) also converges. Such extremum estimators are called M-estimator (Amemiya 1985).

It is well known that $A \overset{P}{\to} B$ (i.e, A is consistent estimator of B) whenever

1. $\mathcal{R}$, the parameter space for F_{N_2}, is a compact set;
2. $\sup_{g \in \mathcal{R}} |F_M(g) - F(g)| \overset{P}{\to} 0$; and
3. $F\left(B\right) < F\left(g\right)$ for any $g \in \mathcal{R}$ such that $g \neq B$,

by the M-estimator consistency (Amemiya 1985). Since we assumed that the function class $\mathcal{R}$ is compact, we only need to check the second and third conditions.

Consider $|F_M(g) - F(g)|$, which is given by

$$\begin{aligned}
&|F_M(g) - F(g)| \\
&= |\sum_{i=1}^{N_1} \widehat{\mathcal{W}}_{1,(i)} \left(\widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right) - \widehat{g_1}\left(\mathbf{X}_{(i)}\right)\right)^2 - \sum_{i=1}^{N_1} \widehat{\mathcal{W}}_{1,(i)} \left(\mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right) - \widehat{g_1}\left(\mathbf{X}_{(i)}\right)\right)^2| \\
&= |\sum_{i=1}^{N_1} \widehat{\mathcal{W}}_{1,(i)} \left(\left(\widehat{r} - g_i\right)^2 - \left(r - g_i\right)^2\right)| \\
&= |\sum_{i=1}^{N_1} \widehat{\mathcal{W}}_{1,(i)} \left(\widehat{r}^2 - r^2 + 2\left(r - \widehat{r}\right)g_i\right)|,
\end{aligned}$$

where $r = \mathcal{B}_2\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)$; $\widehat{r} = \widehat{\mathcal{B}}_{2,N_2}\left(\mathbf{z}, h_y\left(\mathbf{y}\right)\right)$; $g_i = \widehat{g_1}\left(\mathbf{X}_{(i)}\right)$. Note $\widehat{r}$ converges to r for any g by the assumption that $\widehat{r}$ is a consistent estimator of r. This leads that $\widehat{r}^2 \overset{P}{\to} r^2$ too by the continuous mapping theorem. Therefore, we can witness that $\sup_{g \in \mathcal{R}} |F_M(g) - F(g)| \overset{P}{\to} 0$ holds.

Note that the third condition holds by the definition of the quantity B. Therefore, $A \overset{P}{\to} B$, which leads that the composition operators are consistent given assumptions. $\square$.

E Details of Simulations in the Main Paper

Note that $\texttt{N}\left(\mu, \sigma\right)$ represents the Normal distribution with its mean μ and standard deviation σ. Also $\texttt{B}\left(p\right)$ denote the Binomial distribution with the mean p. For a general variable W, $W = \texttt{B}\left(p\right)$ represents that W is generated from $\texttt{B}\left(p\right)$. Also, $\texttt{Sigm}\left(\cdot\right)$ denotes the sigmoid function mapping values to $[0, 1]$. Given two vectors $\mathbf{a} = \{a_1, \cdots, a_D\}$ and $\mathbf{b} = \{b_1, \cdots, b_D\}$, $\mathbf{a}^\intercal \mathbf{b} \equiv \sum_i a_i b_i$. Finally, $[D] \equiv \{1, 2, \cdots, D\}$.

E.1 SCM for Frontdoor (Fig. 2b)

We set X to be binary, Y a continuous variable in $[0,1]$, and $\mathbf{Z} = (Z_1, \ldots, Z_D)$ consisting of D binary variables Z_i. A structural causal model is given as follows:

$$\begin{aligned} U_{X,Y} &\sim \mathtt{N}\left(-2,1\right) \\ f_X\left(U_{X,Y}\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(U_{X,Y} + \epsilon_x\right)\right) \\ f_{Z_i}\left(X\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(c_{1,i} + c_{2,i}X + \epsilon_{z,i}\right)\right) \text{ for all } i \in [D] \\ f_Y\left(U_{X,Y}, \mathbf{Z}\right) &= \mathtt{Sigm}\left(2\mathbf{Z}^\intercal \mathbf{c_y} + U_{X,Y} + \epsilon_y\right) \end{aligned}$$

where $\epsilon_x \sim \mathtt{N}\left(0, 0.5\right)$, $\epsilon_y \sim \mathtt{N}\left(0,1\right)$ and $\epsilon_{z_i} \sim \mathtt{N}\left(-1,1\right)$ for all $i \in [D]$. Also, the parameters $\mathbf{c_1} = \{c_{1,i}\}_{i=1}^{D}$ and $\mathbf{c_2} = \{c_{2,i}\}_{i=1}^{D}$ are given by

$$\begin{aligned} \mathbf{c_1} &\sim \mathtt{N}\left(-2,1\right) \\ \mathbf{c_2} &\sim \mathtt{N}\left(-2,1\right) \\ \mathbf{c_y} &\sim \mathtt{N}\left(1,1\right). \end{aligned}$$

E.2 SCM for mSBD (Fig. 2a)

We set X_1, X_2, Y_1 to be singleton binary variables; Y_2 a continuous variable in $[0,1]$; $Z_i = (Z_{i1}, \ldots, Z_{iD})$ for $i = 1, 2$, where Z_{ij} are binary. A structural causal model is given as follows:

$$\begin{aligned} f_{Z_{1,i}}\left(\cdot\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(c_{a1,i} + c_{a2,i}\epsilon_z + \epsilon_{z_1}\right)\right) \text{ for all } i \in [D] \\ f_{X_1}\left(\mathbf{Z}_1\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(0.2\mathbf{Z_1}^\intercal \mathbf{c_{x_1}} + \epsilon_{x_1}\right)\right) \\ f_{Y_1}\left(\mathbf{Z}_1, X_1\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(\mathbf{Z_1}^\intercal \mathbf{c_{y_1}} - \left(2X_1 - 1\right) - \epsilon_{y_1} - 2\right)\right) \\ f_{Z_{2,i}}\left(\mathbf{Z}_1, X_1, Y_1\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(Z_{1,i}c_{b1,i} + \left(2X_1 - 1\right) + \epsilon_{z_2} + Y_1 - c_{b2,i}\right)\right) \text{ for all } i \in [D] \\ f_{X_2}\left(\mathbf{Z}_1, X_1, Y_1, \mathbf{Z}_2\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(-\mathbf{Z_1}^\intercal \mathbf{c}_{xb1} + \mathbf{Z_2}^\intercal \mathbf{c}_{xb2} + 2X_1 - 1 + Y_1 + 2X_2 - 1 + \epsilon_{x_2}\right)\right) \\ f_{Y_2}\left(\mathbf{Z}_1, X_1, Y_1, \mathbf{Z}_2, X_2\right) &= \mathtt{Sigm}\left(-0.5\left(\mathbf{Z_1}^\intercal \mathbf{c}_{yb1} + \mathbf{Z_2}^\intercal \mathbf{c}_{yb2} + \epsilon_{y_2} - 1\right) + 2X_1 + 2X_2 - 2 - \epsilon_{y_2} + Y_1 - \epsilon_{y_2}\right). \end{aligned}$$

where $\epsilon_{z_1} \sim \mathtt{N}\left(1, 0.5\right)$, $\epsilon_{x_1} \sim \mathtt{N}\left(0,1\right)$, $\epsilon_{y_1} \sim \mathtt{N}\left(0,1\right)$, $\epsilon_{z_2} \sim \mathtt{N}\left(0,0.5\right)$, $\epsilon_{x_2} \sim \mathtt{N}\left(0,1\right)$ and $\epsilon_{y_2} \sim \mathtt{N}\left(0, 0.5\right)$; and the parameters are generated as follow:

$$\begin{aligned} \mathbf{c}_{a1} \sim \mathtt{N}\left(2,1\right)\ , \mathbf{c}_{a2} \sim \mathtt{N}\left(-2,1\right)\ , \mathbf{c_{x_1}} \sim \mathtt{N}\left(-1,0.5\right) \\ \mathbf{c_{y_1}} \sim \mathtt{N}\left(1,0.8\right)\ , \mathbf{c_{b1}} \sim \mathtt{N}\left(1,0.5\right)\ , \mathbf{c_{b2}} \sim \mathtt{N}\left(-1,1\right) \\ \mathbf{c}_{xb1} \sim \mathtt{N}\left(0.3,1\right)\ , \mathbf{c}_{xb2} \sim \mathtt{N}\left(0.2,1\right)\ , \mathbf{c}_{yb1} \sim \mathtt{N}\left(0.3,1\right)\ , \mathbf{c}_{yb2} \sim \mathtt{N}\left(-0.5,1\right); \end{aligned}$$

E.3 SCM for Surrogate endpoints (Fig. 1a)

We set W_2, X to be binary, $\mathbf{W}_1 = \{W_{1,1}, \cdots, W_{1,D}\}$ where each $W_{1,i}$ is binary, and $Y \in [0,1]$. The SCM is given as follows:

$$\begin{aligned} U_1 &\sim \mathtt{N}\left(0,1\right) \\ U_2 &\sim \mathtt{N}\left(0,1\right) \\ f_{W_{1,i}}\left(U_1, U_2\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(c_{1,i} + U1 + c_{2,i} + U2 + \epsilon_{w_1}\right)\right) \\ f_{W_2}\left(\mathbf{W}_1\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(-\left(2\mathbf{W_1} - 1\right)^\intercal \mathbf{c}_{w_2} + \epsilon_{w_2}\right)\right) \\ f_X\left(U_1, W_2\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(U_1 - 4W_2 + 2 + \epsilon_x\right)\right) \\ f_Y\left(U_2, X\right) &= \mathtt{Sigm}\left(0.5U_2 - 2X + 1 + \epsilon_y\right). \end{aligned}$$

where $\left(\epsilon_{w_1}, \epsilon_{w_2}, \epsilon_x, \epsilon_y\right) \sim \mathtt{N}\left(0, 0.5\right)$; $\mathbf{c}_1 \sim \mathtt{N}\left(-2,1\right), \mathbf{c}_2 \sim \mathtt{N}\left(2,1\right)$; and $\mathbf{c}_{w_2} \sim \mathtt{N}\left(1,1\right)$.

E.4 SCM for Causal mediators (Fig. 1b)

We set X to be binary; Y a continuous variable in $[0,1]$; $Z_4 = (Z_{4,1}, \ldots, Z_{4,D})$ where $Z_{4,j}$ are binary; and Z_1, Z_2, Z_3 to be singleton binary variables. A structural causal model is given as follow:

$$\begin{aligned} U_{X,Y} &\sim \mathtt{N}\left(0,2\right) \\ f_{Z_1}\left(\cdot\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(\epsilon_{z_1}\right)\right) \\ f_{Z_2}\left(Z_1\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(-0.5\left(2Z_1 - 1\right) + \epsilon_{z_2}\right)\right) \\ f_{Z_3}\left(Z_1\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(2Z_1 - 1\right) + \epsilon_{z_3}\right) \\ f_X\left(Z_1, Z_2, U_{X,Y}\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(\left(2Z_1 - 1\right)U_{X,Y} - \left(2Z_2 - 1\right)U_{X,Y} + \epsilon_x\right)\right) \\ f_{Z_{4,i}}\left(Z_2, Z_3, X\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(c_{1,i}Z_2X - c_{2,i}Z_3X + \epsilon_{Z_4}\right)\right) \text{ for all } i \in [D] \\ f_Y\left(Z_1, Z_3, Z_4, U_{X,Y}\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(-0.5(\mathbf{2Z_4} - \mathbf{1})^\intercal \mathbf{c}_y + 2Z_1 + 2Z_3 - 2 + 1.5U_{X,Y} + \epsilon_y\right)\right) \end{aligned}$$

where $\epsilon_{z_1} \sim \texttt{N}(0,1)$, $\epsilon_{z_2} \sim \texttt{N}(0,1)$, $\epsilon_{z_3} \sim \texttt{N}(0,1)$, $\epsilon_{z_4} \sim \texttt{N}(0,1)$, $\epsilon_x \sim \texttt{N}(0,1)$, $\epsilon_y \sim \texttt{N}(0,1)$ and

$$\mathbf{c}_1 \sim \texttt{N}(-0.8, 2)\,,\ \mathbf{c}_2 \sim \texttt{N}(1.2, 1)\,, \mathbf{c}_3 \sim \texttt{N}(1, 2)\,.$$

F Details of Simulations in Appendix C.2

F.1 SCM for mSBD (Fig. 2a)

We set X_1, X_2, Y_1 to be singleton binary variables; Y_2 a continuous variable in $[0, 1]$; $Z_i = (Z_{i1}, \ldots, Z_{iD})$ for $i = 1, 2$, where Z_{ij} are binary. A structural causal model is given as follows:

$$\begin{aligned}
f_{Z_{1,i}}(\cdot) &= \texttt{B}\left(\texttt{Sigm}\left(c_{a1,i} + c_{a2,i}\epsilon_z + \epsilon_{z_1}\right)\right) \text{ for all } i \in [D] \\
f_{X_1}(\mathbf{Z}_1) &= \texttt{B}\left(\texttt{Sigm}\left(0.2\mathbf{Z_1}^{\intercal}\mathbf{c_{x_1}}\epsilon_{x_1} + \epsilon_{x_1}\right)\right) \\
f_{Y_1}(\mathbf{Z}_1, X_1) &= \texttt{B}\left(\texttt{Sigm}\left((2\mathbf{Z_1} - 1)^{\intercal}(\mathbf{2c_{y_1}} - \mathbf{1}) - (2X_1 - 1) - \epsilon_{y_1} - 2\right)\right) \\
f_{Z_{2,i}}(\mathbf{Z}_1, X_1, Y_1) &= \texttt{B}\left(\texttt{Sigm}\left(Z_{1,i}c_{b1,i} + (2X_1 - 1) + \epsilon_{z_2} + Y_1 - c_{b2,i}\right)\right) \text{ for all } i \in [D] \\
f_{X_2}(\mathbf{Z}_1, X_1, Y_1, \mathbf{Z}_2) &= \texttt{B}\left(\texttt{Sigm}\left((-(2\mathbf{Z_1} - 1)^{\intercal}(2\mathbf{c}_{xb1} - 1) + (2\mathbf{Z_2} - 1)^{\intercal}(-2\mathbf{c}_{xb2} + 1))\epsilon_{x_2} + 2X_1\epsilon_{x_2} - 1 + Y_1 + 2X_2 - 1 + \epsilon_{x_2}\right)\right) \\
f_{Y_2}(\mathbf{Z}_1, X_1, Y_1, \mathbf{Z}_2, X_2) &= \texttt{Sigm}\left(-0.5\left(\mathbf{Z_1}^{\intercal}\mathbf{c}_{yb1} + \mathbf{Z_2}^{\intercal}\mathbf{c}_{yb2} + \epsilon_{y_2} - 1\right) + 2X_1 + 2X_2 - 2 - \epsilon_{y_2} + Y_1 - \epsilon_{y_2}\right).
\end{aligned}$$

where $\epsilon_{z_1} \sim \texttt{N}(1, 0.5)$, $\epsilon_{x_1} \sim \texttt{N}(0,1)$, $\epsilon_{y_1} \sim \texttt{N}(0,1)$, $\epsilon_{z_2} \sim \texttt{N}(0, 0.5)$, $\epsilon_{x_2} \sim \texttt{N}(0,1)$ and $\epsilon_{y_2} \sim \texttt{N}(0, 0.5)$; and the parameters are generated as follow:

$$\begin{gathered}
\mathbf{c}_{a1} \sim \texttt{N}(2, 1)\ \ , \mathbf{c}_{a2} \sim \texttt{N}(-2, 1)\ \ , \mathbf{c_{x_1}} \sim \texttt{N}(-1, 0.5) \\
\mathbf{c_{y_1}} \sim \texttt{N}(1, 0.8)\ \ , \mathbf{c_{b1}} \sim \texttt{N}(1, 0.5)\ \ , \mathbf{c_{b2}} \sim \texttt{N}(-1, 1) \\
\mathbf{c}_{xb1} \sim \texttt{N}(0.3, 1)\ \ , \mathbf{c}_{xb2} \sim \texttt{N}(0.2, 1)\ \ , \mathbf{c}_{yb1} \sim \texttt{N}(0.3, 1)\ \ , \mathbf{c}_{yb2} \sim \texttt{N}(-0.5, 1)\,;
\end{gathered}$$

F.2 SCM for Surrogate endpoints (Fig. 1a)

We set W_2, X to be binary, $\mathbf{W}_1 = \{W_{1,1}, \cdots, W_{1,D}\}$ where each $W_{1,i}$ is binary, and $Y \in [0, 1]$. The SCM is given as follows:

$$\begin{aligned}
U_1 &\sim \texttt{N}(0, 1) \\
U_2 &\sim \texttt{N}(0, 1) \\
f_{W_{1,i}}(U_1, U_2) &= \texttt{B}\left(\texttt{Sigm}\left(c_{1,i} + U_1 + c_{2,i} + U_2 + \epsilon_{w_1}\right)\right) \\
f_{W_2}(\mathbf{W}_1) &= \texttt{B}\left(\texttt{Sigm}\left(-\left(2\mathbf{W_1} - 1\right)^{\intercal}\mathbf{c}_{w_2} + \epsilon_{w_2}\right)\right) \\
f_X(U_1, W_2) &= \texttt{B}\left(\texttt{Sigm}\left(U_1 - 4W_2 + 2 + \epsilon_x\right)\right) \\
f_Y(U_2, X) &= \texttt{Sigm}\left(0.5U_2 - 2X + 1 + \epsilon_y\right).
\end{aligned}$$

where $(\epsilon_{w_1}, \epsilon_{w_2}, \epsilon_x, \epsilon_y) \sim \texttt{N}(0, 0.5)$; $\mathbf{c}_1 \sim \texttt{N}(-2, 1)$, $\mathbf{c}_2 \sim \texttt{N}(2, 1)$; and $\mathbf{c}_{w_2} \sim \texttt{N}(1, 1)$.

F.3 SCM for Surrogate endpoints with mSBD adjustment (Fig. 1a in Appendix)

A structural causal model is constructed over $(\mathbf{W}_1, W_2, W_3, W_4, W_5, X, Y)$, where W_2, W_3, W_4, W_5, X are all binary variables; Let $\mathbf{W_1} = \{W_{1,1}, \cdots, W_{1,D}\}$ where each $W_{1,i}$ for $i \in [D]$ is binary, and $Y \in [0, 1]$. In particular, the SCM is given as follows:

$$\begin{aligned}
U_{W_1X} &\sim \texttt{N}(0, 3) \\
U_{W_1Y} &\sim \texttt{N}(0, 3) \\
U_{W_1W_3} &\sim \texttt{N}(0, 3) \\
U_{W_3W_2} &\sim \texttt{N}(0, 3) \\
U_{W_3W_5} &\sim \texttt{N}(0, 3) \\
U_{W_5W_4} &\sim \texttt{N}(0, 3) \\
f_{W_{1,i}}(U_{W_1X}, U_{W_1W_3}, U_{W_1Y}) &= \texttt{B}\left(\texttt{Sigm}\left(a_{1,i}U_{W_1X} + a_{2,i}U_{W_1W_3} + a_{3,i}U_{W_1Y} + \epsilon_a\right)\right) \\
f_{W_3}(U_{W_1W_3}, U_{W_3W_5}, U_{W_3W_2}) &= \texttt{B}\left(\texttt{Sigm}\left(b_1U_{W_1W_3} + b_2U_{W_3W_5} + b_3U_{W_3W_2} + \epsilon_b\right)\right) \\
f_{W_2}(\mathbf{W_1}, U_{W_3W_2}) &= \texttt{B}\left(\texttt{Sigm}\left(-0.5(\mathbf{2W_1} - \mathbf{1})^{\intercal}\mathbf{c_{ra}}U_{W_3W_2} + U_{W_3W_2} + \epsilon_r\right)\right) \\
f_{W_4}(W_3, U_{W_5W_4}) &= \texttt{B}\left(\texttt{Sigm}\left(-0.5\left(2W_3 - 1\right)^{\intercal} c_{zb}U_{W_5W_4} + U_{W_5W_4} + \epsilon_z\right)\right) \\
f_{W_5}(U_{W_3W_5}, U_{W_5W_4}) &= \texttt{B}\left(\texttt{Sigm}\left(U_{W_3W_5} + U_{W_5W_4} - 1\right)\right) \\
f_X(U_{W_1X}, W_2, W_4) &= \texttt{B}\left(\text{Normal}\left(Cos(U_{W_1X}(2W_2 - 1)) + \log\left(|U_{W_1X}(2W_4 - 1) + 1|\right)U_{W_1X} + \epsilon_X\right)\right) \\
f_Y(X, U_{W_1Y}) &= \texttt{Sigm}\left(U_{W_1Y} + (2X - 1) - \epsilon_y - 1\right)
\end{aligned}$$

where Normalize $(x\cdot)$ for $x \in \mathbf{x}$ is a mapping $(x - \min(\mathbf{x})) / (\max(\mathbf{x}) - \min(\mathbf{x}))$ the value in $[0, 1]$. Parameters are generated by the following procedure:

$$
\begin{aligned}
\epsilon_{w_1}, \epsilon_{w_3}, \epsilon_{w_4}, \epsilon_x, \epsilon_y &\sim \texttt{N}(0, 0.5) \\
\epsilon_{w_2} &\sim \texttt{N}(-1, 1) \\
\mathbf{a}_1 &\sim \texttt{N}(1, 0.5) \\
\mathbf{a}_2 &\sim \texttt{N}(-1, 0.5) \\
\mathbf{a}_3 &\sim \texttt{N}(0.5, 0.5) \\
b_1 &\sim \texttt{N}(-1, 0.5) \\
b_2 &\sim \texttt{N}(0.5, 0.5) \\
b_3 &\sim \texttt{N}(15, 0.5) \\
\mathbf{c_{ra}} &\sim \texttt{N}(2, 0.5) \\
c_{zb} &\sim \texttt{N}(-2, 0.5).
\end{aligned}
$$

F.4 SCM for Front-door (Fig. 2b)

We set X to be binary, Y a continuous variable in $[0, 1]$, and $\mathbf{Z} = (Z_1, \ldots, Z_D)$ consisting of D binary variables Z_i. A structural causal model is given as follows:

$$
\begin{aligned}
U_{X,Y} &\sim \texttt{N}(0, 2) \\
f_X(U_{X,Y}) &= \texttt{B}(\texttt{Sigm}(2U_{X,Y} + \epsilon_x - 3)) \\
f_{Z_i}(X) &= \texttt{B}(\texttt{Sigm}(c_{1,i}(2X - 1) + c_{2,i}(2X - 1) + \epsilon_{z,i})) \text{ for all } i \in [D] \\
Y_1 &= \texttt{Sigm}(-(2\mathbf{Z} - 1)^\intercal \mathbf{c_y}) \\
Y_2 &= \texttt{Sigm}(\sin((2\mathbf{Z} - 1)U_{X,Y}) + \cos((2\mathbf{Z} - 1)U_{X,Y})) \\
f_Y(U_{X,Y}, \mathbf{Z}) &= \texttt{Sigm}(-Y_1 U_{X,Y} + U \log(|\mathbf{Z}^\intercal \mathbf{1} + Y_1|) Y_2 + 3\cos(Y_2 Y_1))
\end{aligned}
$$

where $\epsilon_x \sim \texttt{N}(0, 0.5)$, $\epsilon_y \sim \texttt{N}(0, 1)$ and $\epsilon_{z_i} \sim \texttt{N}(-1, 1)$ for all $i \in [D]$. Also, the parameters $\mathbf{c_1} = \{c_{1,i}\}_{i=1}^D$ and $\mathbf{c_2} = \{c_{2,i}\}_{i=1}^D$ are given by

$$
\begin{aligned}
\mathbf{c_1} &\sim \texttt{N}(-0.8, 2) \\
\mathbf{c_2} &\sim \texttt{N}(1.2, 1) \\
\mathbf{c_y} &\sim \texttt{N}(1, 2).
\end{aligned}
$$

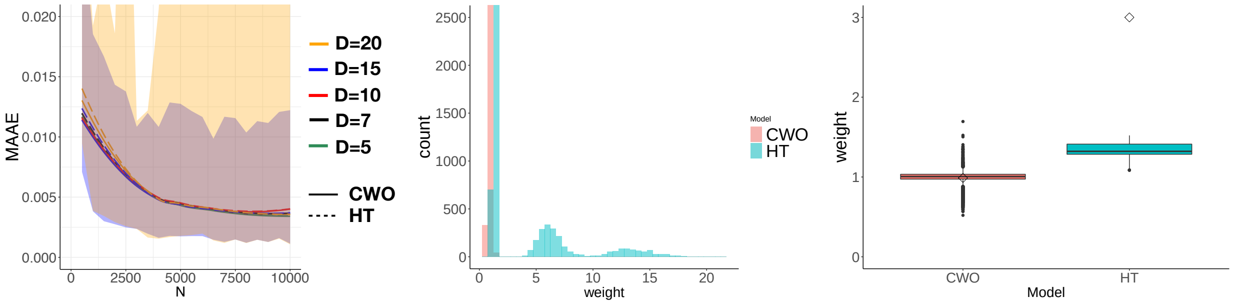
F.5 SCM for Causal mediators (Fig. 1b)

We set X to be binary; Y a continuous variable in $[0, 1]$; $Z_4 = (Z_{4,1}, \ldots, Z_{4,D})$ where $Z_{4,j}$ are binary; and Z_1, Z_2, Z_3 to be singleton binary variables. A structural causal model is given as follow:

$$
\begin{aligned}
U_{X,Y} &\sim \texttt{N}(0, 2) \\
f_{Z_1}(\cdot) &= \texttt{B}(\texttt{Sigm}(\epsilon_{z_1})) \\
f_{Z_2}(Z_1) &= \texttt{B}(\texttt{Sigm}(-0.5(2Z_1 - 1) + \epsilon_{z_2})) \\
f_{Z_3}(Z_1) &= \texttt{B}(\texttt{Sigm}(2Z_1 - 1) + \epsilon_{z_3}) \\
f_X(Z_1, Z_2, U_{X,Y}) &= \texttt{B}(\texttt{Sigm}((2Z_1 - 1) U_{X,Y} - (2Z_2 - 1) U_{X,Y} + \epsilon_x)) \\
f_{Z_{4,i}}(Z_2, Z_3, X) &= \texttt{B}(\texttt{Sigm}(c_{1,i} Z_2 X - c_{2,i} Z_3 X + \epsilon_{Z_4})) \text{ for all } i \in [D] \\
f_Y(Z_1, Z_3, Z_4, U_{X,Y}) &= \texttt{B}(\texttt{Sigm}(-0.5(\mathbf{2Z_4} - \mathbf{1})^\intercal \mathbf{c}_y + 2Z_1 + 2Z_3 - 2 + 1.5U_{X,Y} + \epsilon_y))
\end{aligned}
$$

where $\epsilon_{z_1} \sim \texttt{N}(0, 1)$, $\epsilon_{z_2} \sim \texttt{N}(0, 1)$, $\epsilon_{z_3} \sim \texttt{N}(0, 1)$, $\epsilon_{z_4} \sim \texttt{N}(0, 1)$, $\epsilon_x \sim \texttt{N}(0, 1)$, $\epsilon_y \sim \texttt{N}(0, 1)$ and

$$
\mathbf{c}_1 \sim \texttt{N}(-0.8, 2), \ \mathbf{c}_2 \sim \texttt{N}(1.2, 1), \mathbf{c}_3 \sim \texttt{N}(1, 2).
$$



(a) mSBD (Fig. 2a)

(b) Weights - Histogram

(c) Weights - Box plot

Figure 3: **(a)** MAAE plots (and the 90% confidence interval) comparing the proposed method with the Horvitz-Thompson estimator (HT) for mSBD (Fig. 2a). Confidence intervals for the CWO and HT are shaded in blue and orage, respectivelly. An overlapped area is shaded in pink. **(b,c)** Comparison of variances of weights (when $N = 10000$) estimated through the proposed (CWO) and the H-T method. A diamond dot in (c) denotes the mean estimates of the weights. Plots are rendered in high resolution and can be zoomed in. Best viewed in color.

F.6 SCM for Combinations of Surrogates and Mediators (Fig. 1b in Appendix)

We set W_2, X to be binary, $\mathbf{W}_1 = \{W_{1,1}, \cdots, W_{1,D}\}$ where each $W_{1,i}$ is binary, and $Y \in [0,1]$. The SCM is given as follows:

$$\begin{aligned}
U_1 &\sim \mathtt{N}(-1.5, 2) \\
U_2 &\sim \mathtt{N}(1, 1.5) \\
U_3 &\sim \mathtt{N}(1, 1.2) \\
f_{W_{1,i}}(U_1, U_2) &= \mathtt{B}\left(\mathtt{Sigm}\left(c_{1,i}U_1 + c_{2,i}U_2 + \epsilon_{w_1}\right)\right) \\
f_{W_2}(\mathbf{W}_1) &= \mathtt{B}\left(\mathtt{Sigm}\left(-0.5\left(\mathbf{W_1}\epsilon_{\mathbf{w_2}}\right)^{\intercal}\mathbf{c}_{w_2} + \epsilon_{w_2}\right)\right) \\
f_X(U_1, U_3, W_2) &= \mathtt{B}\left(\mathtt{Sigm}\left(-0.5U_1 - 0.5W_2 + U_3 + \epsilon_x\right)\right) \\
f_{Z_i}(X, U_2) &= \mathtt{B}\left(\mathtt{Sigm}\left(\mathbf{c}_{3,i} + \mathbf{c}_{4,i} + (2X - 1) - U_2 + \epsilon_z\right)\right) \\
f_Y(U_3, Z) &= \mathtt{Sigm}\left(\mathbf{Z}^{\intercal}\mathbf{c}_y + U_3\right).
\end{aligned}$$

where $(\epsilon_{w_1}, \epsilon_{w_2}, \epsilon_x, \epsilon_y) \sim \mathtt{N}(0, 0.5)$; $\mathbf{c}_1 \sim \mathtt{N}(-2, 1)$, $\mathbf{c}_2 \sim \mathtt{N}(2, 1)$; and $\mathbf{c}_{w_2} \sim \mathtt{N}(1, 1)$.

G Comparison with the Horvitz-Thompson Estimator

In this section, we test the proposed method against the Horvitz-Thomson (H-T) estimator. Since the H-T estimators are developed for estimating the causal effect on the BD/SBD settings, the proposed estimator is compared with the H-T estimator only on the mSBD estimands.

G.1 Simulation Setup

We estimate $\mu(\mathbf{x})$ by computing the mean of Y in $D_{int}(\mathbf{v})$, which is treated as the ground truth. We compare the proposed estimators with a parametric plug-in procedure, as discussed next:
Horvitz-Thomson Estimator Another commonly used method in back-door settings is the Horvitz-Thompson (H-T) estimator (Horvitz and Thompson 1952) as an IPW estimator. Given that the causal effect $\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[h(\mathbf{Y})]$ on the mSBD setting could be written as

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[h(\mathbf{Y})] = \mathbb{E}[h(\mathbf{Y})\mathcal{W}_{HT}I_{\mathbf{x}}] \tag{15}$$

where $\mathcal{W}_{HT} \equiv \frac{1}{\prod_i P\left(X_i|\mathbf{X}^{(i-1)}, \mathbf{Y}^{(i-1)}, \mathbf{Z}^{(i)}\right)}$. It is obvious that the estimator for the weight $\mathcal{W}_{HT}$ has higher variance than the weight $\mathcal{W} \equiv \frac{P(\mathbf{x})}{\prod_i P\left(X_i|\mathbf{X}^{(i-1)}, \mathbf{Y}^{(i-1)}, \mathbf{Z}^{(i)}\right)}$ used in the proposed method. For robustly estimating the estimand, using the stabilized weights $\mathcal{W}$ has been recommended in literature (Hernán *et al.* 2002; Karim *et al.* 2017).
Simulated Instances We set X_1, X_2, Y_1 to be singleton binary variables; Y_2 a continuous variable in $[0, 1]$; $Z_i = (Z_{i1}, \ldots, Z_{iD})$ for $i = 1, 2$, where Z_{ij} are binary. A structural causal model is designed in a way that $\prod_i P\left(X_i|\mathbf{X}^{(i-1)}, \mathbf{Y}^{(i-1)}, \mathbf{Z}^{(i)}\right)$ has a small value to explicitly show the robustness of the proposed method compared with the H-T estimator. A causal model is given as follow:

$$
\begin{aligned}
f_{Z_{1,i}}\left(\cdot\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(c_{a1,i} + c_{a2,i}\epsilon_z + \epsilon_{z_1}\right)\right) \text{ for all } i \in [D] \\
f_{X_1}\left(\mathbf{Z}_1\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(2\mathbf{Z_1}^{\intercal}\mathbf{c_{x_1}}\epsilon_{x_1} + \epsilon_{x_1} + 3\right)\right) \\
f_{Y_1}\left(\mathbf{Z}_1, X_1\right) &= \mathtt{B}\left(\mathtt{Sigm}\left((2\mathbf{Z_1} - 1)^{\intercal}(\mathbf{2c_{y_1} - 1}) - (2X_1 - 1) - \epsilon_{y_1} - 2\right)\right) \\
f_{Z_{2,i}}\left(\mathbf{Z}_1, X_1, Y_1\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(Z_{1,i}c_{b1,i} + (2X_1 - 1) + \epsilon_{z_2} + Y_1 - c_{b2,i}\right)\right) \text{ for all } i \in [D] \\
f_{X_2}\left(\mathbf{Z}_1, X_1, Y_1, \mathbf{Z}_2\right) &= \mathtt{B}\left(\mathtt{Sigm}\left(\mathtt{Sigm}\left((-(2\mathbf{Z_1} - 1)^{\intercal}(2\mathbf{c}_{xb1} - 1) + (2\mathbf{Z_2} - 1)^{\intercal}(-2\mathbf{c}_{xb2} + 1)\right)\right)\epsilon_{x_2} + 2X_1(3\epsilon_{x_2} + 1) + Y_1 - 9 + \epsilon_{x_2}\right)\right) \\
f_{Y_2}\left(\mathbf{Z}_1, X_1, Y_1, \mathbf{Z}_2, X_2\right) &= \mathtt{Sigm}\left(-0.5\left(\mathbf{Z_1}^{\intercal}\mathbf{c}_{yb1} + \mathbf{Z_2}^{\intercal}\mathbf{c}_{yb2} + \epsilon_{y_2} - 1\right) + 2X_1 + 2X_2 - 2 - \epsilon_{y_2} + Y_1 - \epsilon_{y_2}\right).
\end{aligned}
$$

where $\epsilon_{z_1} \sim \mathtt{N}\left(1, 0.5\right)$, $\epsilon_{x_1} \sim \mathtt{N}\left(0, 1\right)$, $\epsilon_{y_1} \sim \mathtt{N}\left(0, 1\right)$, $\epsilon_{z_2} \sim \mathtt{N}\left(0, 0.5\right)$, $\epsilon_{x_2} \sim \mathtt{N}\left(0, 1\right)$ and $\epsilon_{y_2} \sim \mathtt{N}\left(0, 0.5\right)$; and the parameters are generated as follow:

$$
\begin{aligned}
\mathbf{c}_{a1} \sim \mathtt{N}\left(2, 1\right)\ , \mathbf{c}_{a2} \sim \mathtt{N}\left(-2, 1\right)\ , \mathbf{c_{x_1}} \sim \mathtt{N}\left(-1, 0.5\right) \\
\mathbf{c_{y_1}} \sim \mathtt{N}\left(1, 0.8\right)\ , \mathbf{c_{b1}} \sim \mathtt{N}\left(1, 0.5\right)\ , \mathbf{c_{b2}} \sim \mathtt{N}\left(-1, 1\right) \\
\mathbf{c}_{xb1} \sim \mathtt{N}\left(0.3, 1\right)\ , \mathbf{c}_{xb2} \sim \mathtt{N}\left(0.2, 1\right)\ , \mathbf{c}_{yb1} \sim \mathtt{N}\left(0.3, 1\right)\ , \mathbf{c}_{yb2} \sim \mathtt{N}\left(-0.5, 1\right);
\end{aligned}
$$

G.2 Simulation Results

We test the proposed estimator (CWO) against the H-T estimator on the mSBD estimand. A description about the simulation instance is given in Appendix F.

We test on estimating $\mathbb{E}_{P_{x_1,x_2}(y_2)}\left[Y_2\right]$, for which the CWO and H-T estimators use results from Prop. 1 and Thm. 2 respectively. We set X_1, X_2, Y_1 to be binary, Y_2 continuous within $[0, 1]$, and $Z_i = (Z_{i1}, \ldots, Z_{iD})$ for $i = 1, 2$, where all Z_{ij} are binary. Figure 3a presents the MAAE plots for $D = 5, 7, 10, 15, 20$. To explicitly represent the uncertainty of the estimator, we represent the 90% confidence intervals by shading the area between 5th percentiles and 95th percentiles of the MAAE curve with $D = 20$ in blue (CWO) and red (H-T). The overlapped area is shaded in pink. For visual clarity, we only include the confidence interval for $D = 20$ in Fig. 3a. The MAAE plot shows that the CWO performs at least on par with the H-T estimator in average. As one could see, however, the resultant estimates of the H-T estimator is less reliable due to its high variances, as illustrated using the confidence intervals.

As expected, the proposed estimator provides more robust estimates compared to the H-T estimator. Specifically, Fig. 3a shows that the resultant estimates of the proposed method has smaller variances than the results of the H-T estimator. We observe that the estimates weights of the proposed method ($\mathcal{W}$) is more robust compared to the weight ($\mathcal{W}_{HT}$). Fig. 3(b,c) shows that the estimates of the weights of the proposed method has smaller variances, concentrated on the mean estimates.

References

Takeshi Amemiya. *Advanced econometrics*. Harvard university press, 1985.

Peter C Austin and Ewout W Steyerberg. Events per variable and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical methods in medical research*, 26(2):796–808, 2017.

Isabel R Fulcher, Ilya Shpitser, Stella Marealle, and Eric J Tchetgen Tchetgen. Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2019.

Miguel A Hernan and James M Robins. *Causal Inference*. Boca Raton: Chapman & Hall/CRC, 2019. forthcoming.

Miguel A Hernán, Babette A Brumback, and James M Robins. Estimating the causal effect of zidovudine on cd4 count with a marginal structural model for repeated measures. *Statistics in medicine*, 21(12):1689–1709, 2002.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Mohammad Ehsanul Karim, John Petkau, Paul Gustafson, Helen Tremlett, and The Beams Study Group. On the application of statistical learning approaches to construct inverse probability weights in marginal structural cox models: hedging against weight-model misspecification. *Communications in Statistics-Simulation and Computation*, 46(10):7668–7697, 2017.

Bryan Lim, Ahmed Alaa, and Mihaela van der Schaar. Forecasting treatment responses over time using recurrent marginal structural networks. In *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems*, pages 7483–7493, 2018.

Henry B Mann and Abraham Wald. On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14(3):217–226, 1943.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. 2nd edition, 2009.

Peter Peduzzi, John Concato, Elizabeth Kemper, Theodore R Holford, and Alvan R Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 49(12):1373–1379, 1996.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 2000.

James M Robins. Marginal structural models. In *1997 Proceedings of the American Statistics Associations*. American Statistics Associations, 1997.

James M Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer, 2000.

Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the 18th conference on Uncertainty in artificial intelligence*, pages 519–527. Morgan Kaufmann Publishers Inc., 2002.

Jin Tian. *Studies in Causal Reasoning and Learning*. PhD thesis, Computer Science Department, University of California, Los Angeles, CA, November 2002.

Tyler J VanderWeele. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1):18–26, 2009.

Eric Vittinghoff and Charles E McCulloch. Relaxing the rule of ten events per variable in logistic and cox regression. *American journal of epidemiology*, 165(6):710–718, 2007.

Junfeng Wen, Negar Hassanpour, and Russell Greiner. Weighted gaussian process for estimating treatment effect. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, 2018.