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 statistical 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:

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

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 $\mathbb{E}[Y|x, z]$ with some statistical significance is hardly realizable.

A series of robust and efficient statistical estimators for estimating the BD estimand (Eq. (1)) from finite samples have been developed to circumvent these challenges with great practical success, including the 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 has 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 experi-

mental 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. 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 level 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. □

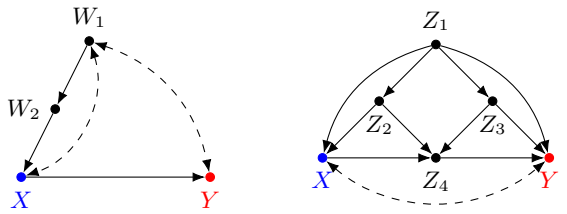
In general, many graphical and algorithmic conditions have been developed for determining the identifiability of a target causal effect $P_x(y)$ in a given causal graph. However, no general method exists in the literature for estimating $P_x(y)$ from finite samples whenever it is identifiable (for example 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 the 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, our goal is to start bridging this gap in the literature between causal "identification" and causal "estimation". Specifically, we propose to extend weighting-based methods developed for BD estimand (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 introduce novel machinery for estimating complex causal effects - composition of weighting operators.

3. We develop 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.

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



(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.

The proofs are provided in Appendix D in suppl. 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 $\widehat{P}(\mathbf{v})$ 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

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 $D^{\mathcal{W}}_{obs} \sim P^{\mathcal{W}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ represent the reweighted D_{obs}, Prop. 1 says $D^{\mathcal{W}}_{obs}$ 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 $X_1, \ldots, X_n$ and corresponding covariates $Z_1, \ldots, Z_n$. 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

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

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*

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

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

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

$$\mathbb{E}_{P_{x_1,x_2}(y_2)}[Y_2] = \mathcal{B}[Y_2 \mid \{x_1, x_2\}; \mathcal{W}], \quad (8)$$
$$\text{where } \mathcal{W} = \frac{P(x_1, x_2)}{P(x_1|z_1)P(x_2|x_1, y_1, z_1, z_2)}.$$

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) $(\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{R})_{G_{\underline{\mathbf{X}}\overline{\mathbf{R}}}}$; and (3) $\mathbf{Z}$ is mSBD admissible relative to $(\mathbf{R}, (\mathbf{X}, \mathbf{Y}))$.

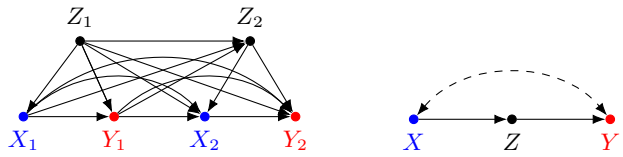
Theorem 3. *If $(\mathbf{R}, \mathbf{Z})$ is surrogate admissible relative to $(\mathbf{X}, \mathbf{Y})$, then*[4]

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

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

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

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



(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

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

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

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

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

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

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*

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

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

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

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

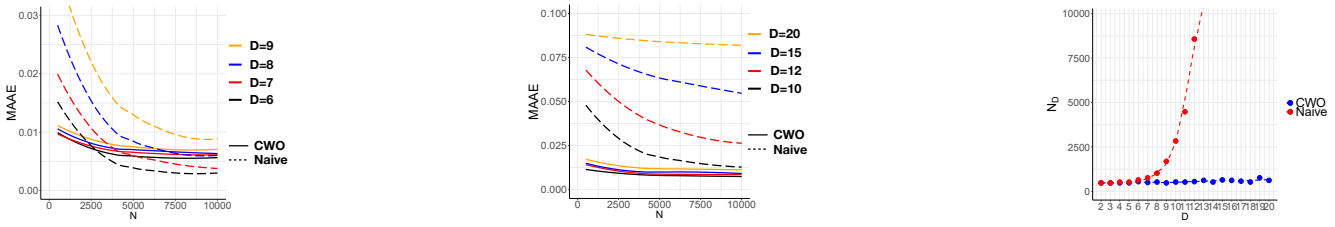
Finally, we obtain that the expected causal effect $\mathbb{E}_{P_x(y)}[Y] = \mathbb{E}_{P_x(\mathbf{z})}\left[\mathbb{E}_{P_{\mathbf{z}}(y)}[Y]\right]$ is given by $(\mathcal{B}_1 \circ \mathcal{B}_2)(x)$.

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 $\widehat{\mathcal{B}}[h(\mathbf{Y}) \mid \mathbf{x}; \mathcal{W}](\mathbf{x}) \equiv g^*(\mathbf{x})$ is estimated by the weighted regression as follows:

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

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 $\widehat{\mathcal{W}}(\mathbf{V}_{(i)}) = \widehat{P}(\mathbf{x}_{(i)}) / \widehat{P}(\mathbf{x}_{(i)} \mid \mathbf{z}_{(i)})$. 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 $P(\mathbf{x}|\mathbf{z}) = \prod_{d=1}^{n} P(x_d|\mathbf{z}, \mathbf{x}^{(d-1)})$. 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 $\widehat{\mathcal{B}}_1(\mathbf{x}) \equiv \widehat{\mathcal{B}}[h_{\mathbf{z}}(\mathbf{Z}) \mid \mathbf{x}; \mathcal{W}_1]$ and $\widehat{\mathcal{B}}_2(\mathbf{z}) \equiv \widehat{\mathcal{B}}[h_{\mathbf{y}}(\mathbf{Y}) \mid \mathbf{z}; \mathcal{W}_2]$, 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 $\widehat{\mathcal{B}}_1(\mathbf{x}) \equiv \widehat{\mathcal{B}}[h_{\mathbf{z}}(\mathbf{Z}) \mid \mathbf{x}; \mathcal{W}_1]$ and $\widehat{\mathcal{B}}_2(\mathbf{z}) \equiv \widehat{\mathcal{B}}[h_{\mathbf{y}}(\mathbf{Y}) \mid \mathbf{z}; \mathcal{W}_2]$. The empirical composition $\widehat{(\mathcal{B}_1 \circ \mathcal{B}_2)}(\mathbf{x})$ is defined by

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

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

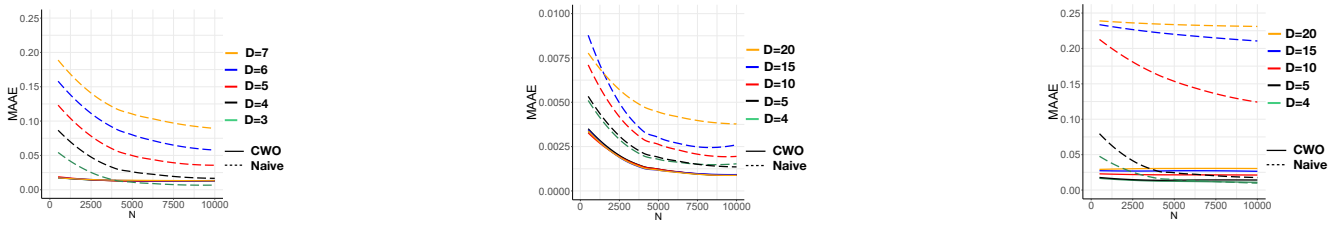
Theorem 6 (Consistency of the composition)**.** *Let $\widehat{\mathcal{B}}_1(\mathbf{x})$ and $\widehat{\mathcal{B}}_2(\mathbf{z})$ be consistent estimators of $\mathcal{B}_1(\mathbf{x})$ and $\mathcal{B}_2(\mathbf{z})$. Let the function class $\mathcal{R}_1$ of $\widehat{\mathcal{B}}_1$ be a compact 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. $\mathbb{E}[Y|x_1, x_2, z_1, z_2, y_1]$ 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 $\mathbb{E}_{P_{\mathbf{x}}(y)}[Y]$ 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 $\mathbb{E}_{P_x(y)}[Y]$ 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 $\mathbb{E}_{P_{x_1,x_2}(y_2)}[Y_2]$ 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 $\mathbb{E}_{P_x(y)}[Y]$ (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 $\mathbb{E}_{P_x(y)}[Y]$. 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 observed distribution 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 case? On the other hand, can weighting operators be further composed to identify effects beyond the decomposability criterion? Further, 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*.