PRELIMINARY VERSION: DO NOT CITE
The AAAI Digital Library will contain the published
version some time after the conference

Estimating Identifiable Causal Effects through Double Machine Learning

Yonghan Jung,[1] **Jin Tian,** [2] **Elias Bareinboim** [3]
[1] Department of Computer Science, Purdue University
[2] Department of Computer Science, Iowa State University
[3] Department of Computer Science, Columbia University
jung222@purdue.edu, jtian@iastate.edu, eb@cs.columbia.edu

Abstract

Identifying causal effects from observational data is a pervasive challenge found throughout the empirical sciences. Very general methods have been developed to decide the identifiability of a causal quantity from a combination of observational data and causal knowledge about the underlying system. In practice, however, there are still challenges to estimating identifiable causal functionals from finite samples. Recently, a method known as *double/debiased machine learning* (DML) (Chernozhukov et al. 2018) has been proposed to learn parameters leveraging modern machine learning techniques, which is both robust to model misspecification and bias-reducing. Still, DML has only been used for causal estimation in settings when the back-door condition (also known as conditional ignorability) holds. In this paper, we develop a new, general class of estimators for *any* identifiable causal functionals that exhibit DML properties, which we name DML-ID. In particular, we introduce a complete identification algorithm that returns an influence function (IF) for any identifiable causal functional. We then construct the DML estimator based on the derived IF. We show that DML-ID estimators hold the key properties of debiasedness and doubly robustness. Simulation results corroborate with the theory.

1 Introduction

Inferring causal effects from observational data is a fundamental task throughout the data-intensive sciences. There exists a growing literature trying to understand the conditions under which causal conclusions can be drawn from non-experimental data, which comes under the rubric of *causal inference* (Pearl 2000; Pearl and Mackenzie 2018). In particular, the literature of *causal effect identification* (Pearl 2000, Def. 3.2.4) investigates the conditions under which an interventional distribution $P(Y = y|do(X = x))$ (for short, $P_x(y)$), representing the causal effect of the treatment X on the outcome Y, could be inferred from the observational distribution $P(V)$ and the causal graph G. Causal effect identification under various settings has been extensively studied, and algorithms and graphical conditions have been developed (Pearl 1995; Tian and Pearl 2003; Huang and Valtorta 2006; Shpitser and Pearl 2006; Bareinboim and Pearl 2012, 2016; Jaber, Zhang, and Bareinboim 2018; Lee, Correa, and Bareinboim 2019, 2020; Lee and Bareinboim 2020).

As a specific example, the celebrated *back-door (BD)* condition (Pearl 2000, Sec. 3.3.1) (known as *ignorability* in statistics (Rubin 1978)) states that $P_x(y)$ could be identified by adjustment – i.e., $P_x(y) = \sum_z P(y|x,z)P(z)$ – whenever there exists a set of covariates Z that blocks all the backdoor paths between X and Y in the causal graph G. Identification algorithms express a target effect in terms of the observational distribution, then one needs to go further, and estimate the resulting expression from finite samples. In practice, whenever the number of samples are finite and the set of covariates (e.g., Z) is high dimensional – i.e., almost always – estimating causal expressions is quite challenging.

Effective estimators have been developed for specific settings. For instance, a plethora of estimators have been developed for the family of BD settings, including point and time-series forms (*Sequential BD*, or SBD) (Pearl and Robins 1995); also called the g-formula (Robins 1986). These estimators include regression-based methods (e.g., (Hill 2011; Shalit, Johansson, and Sontag 2017)) or weighting-based methods (Horvitz and Thompson 1952; Robins, Hernan, and Brumback 2000; Johansson et al. 2018), to name a few. More recently, estimators have been developed for identifiable causal functionals under settings beyond the typical BD/SBD (Jung, Tian, and Bareinboim 2020a,b).

Further, doubly robust estimators have been developed for the BD/SBD setting to address model misspecification (Robins, Rotnitzky, and Zhao 1994; Bang and Robins 2005; Van Der Laan and Rubin 2006; Benkeser et al. 2017; Rotnitzky and Smucler 2019; Smucler, Sapienza, and Rotnitzky 2020), and more recently, for a few specific settings (Fulcher et al. 2019; Bhattacharya, Nabi, and Shpitser 2020).

One noticeable feature shared across the aforementioned estimators is the need of estimating conditional probabilities (e.g., $P(y|x,z)$, $P(z)$), called *nuisance functions*, or *nuisance* in short. Typically nuisance functions are estimated by fitting a parametric model such as logistic regression. In recent years, there is an explosion in the use of modern machine learning (ML) methods to account for very complex and high-dimensional nuisance functions, which include random forests, boosted regression trees, deep neural

networks, to cite some prominent examples. However, these methods inherently use regularization to control overfitting, which often translates into acute bias in estimators of the causal estimands. In practice, this means that these estimators will not be able to achieve $\sqrt{N}$-consistency, where N is the sample size, which is usually desirable.

Recently, a powerful method called *double/debiased machine learning* (DML) (Chernozhukov et al. 2018) has been proposed to provide '*debiased*' estimators, which achieve $\sqrt{N}$-consistency with respect to the target estimand, while admitting the use of a broad array of modern ML methods for estimating the nuisances (including random forests, neural nets, etc). The DML has been developed and applied in the context of causal functional estimation in a few specific settings. (Zadik, Mackey, and Syrgkanis 2018; Syrgkanis et al. 2019; Foster and Syrgkanis 2019; Chernozhukov et al. 2019; Kallus and Uehara 2020; Farbmacher et al. 2020).

Even though there exists a complete framework for estimating arbitrary identifiable causal functionals based on ML (Jung, Tian, and Bareinboim 2020b), the corresponding procedures do not exhibit DML properties. On the other hand, there are effective and robust estimators for the BD case, which is only a fraction of all the identifiable causal functionals. In this paper, we aim to bridge this gap by developing DML estimators for any identifiable causal estimand, moving beyond the BD/ignorability family. For concreteness, consider the following two examples[1].

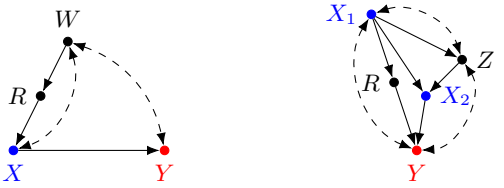
Example 1. A data scientist aims to establish how cardiac output (X) affects the blood pressure (Y) from observational data. In the causal model shown in Fig. 1a, the heart rate (R) directly causes X, while being influenced by the level of catecholamine (W), a hormone released in response to stress. The level of total peripheral resistance (U_1) affects W and X, and the level of the analgesia (U_2) influences W and Y. Both U_1 and U_2 are unobserved confounders due to complications in measurement (left implict as a dashed-bidirected arrow). A standard identification algorithm derives the causal effect $P_x(y)$ as:

$$P_x(y) = (\sum_w P(y, x|r, w)P(w))/(\sum_w P(x|r, w)P(w)). \quad (1)$$

Example 2. Suppose the data scientist needs to establish the effect of a new treatment based on the cardiovascular shunt (X_1) and the lung ventilation (X_2) on catecholamine (Y). In the causal model in Fig. 1b, X_1 directly affects the ventilation tube (Z), the level of arterial oxygen saturation (R), and X_2. Further, Z influences X_2. X_2 and R have direct impact on Y. There are also unmeasured confounders affecting this process: pulmonary embolism (U_1) affects X_1 and Z, the level of total peripheral resistance (U_2) affects X_1 and Y, and the level of the anesthesia (U_3) affects Z and Y. Despite of these unobserved confounders, the effect of interest $P_{x_1,x_2}(y)$ can be identified as

$$P_{x_1,x_2}(y) = \sum_r P(r|x_1) \sum_{x_1',z} P(y|r, x_1', x_2, z)P(z, x_1'). \quad (2)$$

[1]The causal graphs are constructed from the classic 'Alarm' network (Beinlich et al. 1989), originally collected from a system used to monitor patients' conditions.



(a) Example 1 (b) Example 2

Figure 1: Causal graphs corresponding to Examples (1,2). Nodes representing the treatment and outcome are marked in blue and red respectively.

A few observations follow from these two examples. First, note that the estimands of Eqs. (1) or (2) are not in the form of the backdoor adjustment, which means that previous work is not applicable, and no debiased or doubly robust estimators are readily available for such cases. Second, in fact, the only viable method currently available for estimating arbitrary identified causal estimands, beyond a few special settings, is the "plug-in" estimators (Casella and Berger 2002), which estimate nuisance functions and plug them into the equation. However, the plug-in estimators are exposed to the risk of model misspecification since all nuisance functions need to be correctly specified for the estimator to be consistent. Also, they often suffer from the bias caused by the use of flexible ML models in high-dimensional cases under finite samples.

In this paper, we develop DML estimators for any causal effects that is identifiable given a causal graph. More specifically, our contributions are as follows:

1. We develop a systematic procedure for deriving influence functions (IFs) for estimands of any identifiable causal effects.

2. We develop DML estimators for any identifiable causal effect, which enjoy debiasedness and doubly robustness against model misspecification and bias. Experimental studies corroborate our results.

The proofs are provided in Appendix A in suppl. material.

2 Preliminaries

Notations. Each variable is represented with a capital letter (X) and its realized value with the small letter (x). We use bold letters ($\mathbf{X}$) to denote sets of variables. Given an ordered set $\mathbf{X} = (X_1, \cdots, X_n)$ such that $X_i \prec X_j$ for $i < j$, we denote $\mathbf{X}^{(i)} = \{X_1, \cdots, X_i\}$, $\mathbf{X}^{\geq i} = \{X_i, \cdots, X_n\}$, and set $\mathbf{X}^{(i)} = \emptyset$ for $i < 1$. We use $I_{\mathbf{v}'}(\mathbf{V})$ to represent the indicator function such that $I_{\mathbf{v}'}(\mathbf{V}) = 1$ if and only if $\mathbf{V} = \mathbf{v}'$; $I_{\mathbf{v}'}(\mathbf{V}) = 0$ otherwise. We denote $\mathcal{D} = \{\mathbf{V}_{(i)}\}_{i=1}^N$ as samples drawn from $P(\mathbf{V})$, and $\widehat{P}$ the estimated distribution; $\mathbb{E}_P[f(\mathbf{V})]$ denotes the expectation of $f(\mathbf{V})$ over $P(\mathbf{v})$.

We use the typical graph terminology $Pa(\mathbf{C})_G, Ch(\mathbf{C})_G, De(\mathbf{C})_G, An(\mathbf{C})_G$ to represent the union of $\mathbf{C}$ with its parents, children, descendants, ancestors in the graph G. We use $ND(\mathbf{C})$ to denote the nondescendants of any variables in $\mathbf{C}$ (i.e., $ND(\mathbf{C}) \equiv \mathbf{V} \backslash De(\mathbf{C})$).

For a given topological order in G, we use $Pre(\mathbf{C})$ to denote the union of the predecessors of $C_i \in \mathbf{C}$ in G. $G(\mathbf{C})$ denotes the subgraph of G over $\mathbf{C}$. The *latent projection* of a graph G over $\mathbf{V}$ on $\mathbf{C} \subseteq \mathbf{V}$, denoted $G[\mathbf{C}]$, is a graph over $\mathbf{C}$ such that, in addition to edges in $G(\mathbf{C})$, for every pair of vertices $(V_i, V_j) \in \mathbf{C}$, (1) add a directed edge $V_i \to V_j$ in $G[\mathbf{C}]$ if there exists a directed path from V_i to V_j in G such that every vertex on the path is not in $\mathbf{C}$; (2) add a bidirected edge $V_i \leftrightarrow V_j$ in $G[\mathbf{C}]$ if there exists a divergent path between V_i and V_j in G such that every vertex on the path is not in $\mathbf{C}$ (Tian and Pearl 2003). 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 outgoing edges from $\mathbf{C}_2$ in G.

Structural Causal Models. We use the language of structural causal models (SCMs) as our basic semantical framework (Pearl 2000). Each SCM M over a set of variables $\mathbf{V}$ induces a distribution $P(\mathbf{v})$ and a causal graph G, where solid-directed arrows encode functional relationships between observed variables, and dashed-bidirected arrows encode unobserved latent causes (e.g., see Fig. 1a)[2]. 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 interventional distribution $P(\mathbf{v}|do(\mathbf{x})) \equiv P_{\mathbf{x}}(\mathbf{v})$. We refer readers to (Pearl 2000; Bareinboim et al. 2020) for a more detailed discussion of SCMs.

Causal Effect Identification. Given a graph G over $\mathbf{V}$, an effect $P_{\mathbf{x}}(\mathbf{y})$ is *identifiable* in G if $P_{\mathbf{x}}(\mathbf{y})$ is uniquely computable from the observed distribution $P(\mathbf{v})$ in any SCM that induces G (Pearl 2000, p. 77). Complete identification algorithms have been developed based on a decomposition strategy using so-called *confounded components*.

Definition 1 (C**-component (Tian and Pearl 2002)).** In a causal graph, two variables are said to be in the same confounded component (for short, C-component) if and only if they are connected by a bi-directed path, i.e., a path composed solely of bi-directed edges $V_i \leftrightarrow V_j$.

For any $\mathbf{C} \subseteq \mathbf{V}$, the quantity $Q[\mathbf{C}] \equiv P_{\mathbf{v}\setminus\mathbf{c}}(\mathbf{c})$, called a *$C$-factor*, is defined as the post-intervention distribution of $\mathbf{C}$ under an intervention on $\mathbf{V}\setminus\mathbf{C}$. (Tian and Pearl 2003) showed that the causal effect $P_{\mathbf{x}}(\mathbf{y})$ can be represented as a marginalization over a product of C-factors: $P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{d}\setminus\mathbf{y}} Q[\mathbf{D}] = \sum_{\mathbf{d}\setminus\mathbf{y}} \prod_{i=1}^{k_d} Q[\mathbf{D}_i]$, where $\mathbf{D} \equiv An(\mathbf{Y})_{G(\mathbf{V}\setminus\mathbf{X})}$ and $\mathbf{D}_i$ are C-components in $G(\mathbf{D})$.

Semiparametric Theory. Our goal is to estimate an identifiable causal effect $P_{\mathbf{x}}(\mathbf{y})$ from finite samples $\mathcal{D} = \{\mathbf{V}_{(i)}\}_{i=1}^{N}$ drawn from $P(\mathbf{V})$. Assume one aims to estimate a target estimand $\psi \equiv \Psi(P)$ that is a functional of P. For example, $\Psi(P) = \sum_z P(y|x,z)P(z)$. We will leverage the semiparametric theory [3]. Let $P_t \equiv P(\mathbf{v})(1 + tg(\mathbf{v}))$ for any $t \in \mathbb{R}$ and bounded mean-zero random functions $g(\cdot)$ over random variables $\mathbf{V}$, called a *parametric submodel*. If a functional $\Psi(P_t)$ is pathwise (formally, Gâteaux) differentiable at $t = 0$, then there exists a function $\phi(\mathbf{V}; \psi, \eta(P))$ (shortly ϕ), called the *influence function (IF) for the target functional* ψ, where $\eta(P)$ stands for the set of nuisance functions comprising ϕ, satisfying $\mathbb{E}_P[\phi] = 0$, $\mathbb{E}_P[\phi^2] < \infty$, and $\frac{\partial}{\partial t}\Psi(P_t)|_{t=0} = \mathbb{E}_P[\phi(\mathbf{V}; \psi, \eta(P))S_t(\mathbf{V}; t=0)]$ where $S_t(\mathbf{v}; t=0) \equiv \frac{\partial}{\partial t}\log P_t(\mathbf{v})|_{t=0}$ is the score function (Van der Vaart 2000, Chap. 25). An IF ϕ characterizes an estimator T_N satisfying $T_N - \psi = \frac{1}{N}\sum_{i=1}^{N}\phi(\mathbf{V}_{(i)}; \psi, \eta(P)) + o_P(N^{-1/2})$ where $o_P(N^{-1/2})$ is a term that converges in probability with a rate of at least $N^{-1/2}$. Such T_N is a *Regular and Asymptotic Linear* (RAL) estimator of ψ (Van der Vaart 2000, Lemma 25.23). When the IF can be decomposed as $\phi(\mathbf{V}; \psi, \eta(P)) = \mathcal{V}(\mathbf{V}; \eta(P)) - \psi$ for some function $\mathcal{V}(\mathbf{V}; \eta(P))$, called the *uncentered influence function (UIF)*, the corresponding RAL estimator is given by $T_N = \frac{1}{N}\sum_{i=1}^{N}\mathcal{V}(\mathbf{V}_{(i)}, \eta(\widehat{P}))$ (Kennedy 2018).

Double/Debiased Machine Learning (DML). DML methods (Chernozhukov et al. 2018) are based on two ideas: (1) Use a *Neyman orthogonal score*[4] to estimate the target ψ, and (2) Use *cross-fitting* to construct the estimator. Making use of Neyman-orthogonal scores reduces sensitivity with respect to nuisance parameters. Cross-fitting reduces bias induced by overfitting. DML estimators provide $\sqrt{N}$-consistent estimates of the target ψ even when possibly complex or high-dimensional nuisance functions are estimated at slower $N^{-1/4}$ rates ('*debiasedness*') (Chernozhukov et al. 2018). Neyman-orthogonal scores may be constructed using IFs, and under some settings, may coincide with IFs (Chernozhukov et al. 2016).

3 Expressing Causal Effects through a Combination of mSBDs

Our goal is to develop DML estimators for any identifiable causal effects $\psi = P_{\mathbf{x}}(\mathbf{y})$. Towards this goal, we present in this section a sound and complete algorithm that expresses any identifiable causal effects as a combination of *marginalization/multiplication/divisions* (which will be called '*arithmetic combination*') of so-called mSBD estimands. Based on this result, in the subsequent section, we derive an IF

[2]The class of SCMs inducing a directed acyclic graph (DAG) with bidirected arrows is usually called semi-Markovian (Pearl 2000, p. 30). In general, a DAG with arbitrary latent variables can be converted into a DAG with bidirected arrows, i.e. a semi-Markovian model, by computing its latent projection on the set of observed variables. One can show that the projection operation preserves causal identification (Tian and Pearl 2003, Section 6).

[3]The aforementioned causal effect identification theory has been developed under a non-parametric setting, i.e., without any parametric assumptions on the form of the SCM. To estimate an identified estimand $P_{\mathbf{x}}(\mathbf{y}) = \Psi(P)$, imposing strong parametric assumptions over the estimator would go against the non-parametric nature of the identification step. Semiparametric models capture the structural constraints (e.g., conditional independences) imposed by the causal graph while allowing nonparametric models for estimating nuisance functionals (e.g., highly flexible machine learning models such as multi-layered neural networks).

[4]A Neyman orthogonal score is a score function ϕ satisfying $\mathbb{E}_P[\phi(\mathbf{V}; \psi, \eta(P))] = 0$ and $\frac{\partial}{\partial \eta(P_t)}\mathbb{E}_P[\phi(\mathbf{V}; \psi, \eta(P_t))]|_{t=0} = 0$ (Chernozhukov et al. 2016, 2018).

for ψ (that turns out to be a Neyman orthogonal score) by first deriving an IF for mSBD estimands and using them as buildig blocks. We first define the mSBD criterion:

Definition 2 (mSBD criterion (Jung, Tian, and Bareinboim 2020a)). *Given the pair of sets $(\mathbf{X}, \mathbf{Y})$, let $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ be topologically ordered as $X_1 \prec X_2 \prec \cdots \prec X_n$. 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 $i = 1, \cdots, n$. A sequence $\mathbf{Z} = (\mathbf{Z}_1, \cdots, \mathbf{Z}_n)$ is mSBD admissible relative to $(\mathbf{X}, \mathbf{Y})$ if it holds that $\mathbf{Z}_i \subseteq ND\left(\mathbf{X}^{\geq i}\right)$, and $\left(\mathbf{Y}^{\geq i} \perp\!\!\!\perp X_i | \mathbf{Y}^{(i-1)}, \mathbf{Z}^{(i)}, \mathbf{X}^{(i-1)}\right)_{G_{\underline{X_i}\overline{\mathbf{X}^{\geq i+1}}}}$ for $i = 1, \cdots, n$.*

We will use the mSBD criterion as a foundation to construct general causal estimands. To this end, we formally define the notion of a mSBD-operator:

Definition 3 (mSBD operator $\mathcal{M}$). Let $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = ((X_i)_{i=1}^n, (\mathbf{Y}_i)_{i=0}^n, (\mathbf{Z}_i)_{i=1}^n)$ be disjoint sets of ordered variables. The *mSBD operator* $\mathcal{M}[\mathbf{y} \mid \mathbf{x}; \mathbf{z}]$ is defined by

$$\mathcal{M}[\mathbf{y} \mid \mathbf{x}; \mathbf{z}] \equiv \sum_{\mathbf{z}} \prod_{k=0}^{n} P\left(\mathbf{y}_k | \mathbf{x}^{(k)}, \mathbf{z}^{(k)}, \mathbf{y}^{(k-1)}\right) \times \prod_{j=1}^{n} P\left(\mathbf{z}_j | \mathbf{x}^{(j-1)}, \mathbf{z}^{(j-1)}, \mathbf{y}^{(j-1)}\right). \quad (3)$$

If $\mathbf{Z}$ satisfies the mSBD criterion relative to $(\mathbf{X}, \mathbf{Y})$, then the causal effect $P_{\mathbf{x}}(\mathbf{y})$ is identifiable by $P_{\mathbf{x}}(\mathbf{y}) = \mathcal{M}[\mathbf{y} \mid \mathbf{x}; \mathbf{z}]$ (Jung, Tian, and Bareinboim 2020a).

We will develop a systematic procedure that can express causal effects into the arithmetic combinations of mSBD operators. Our algorithm will leverage the existing complete identification procedure in (Tian and Pearl 2003). To establish the connection, we show next how specific C-factors can be identified in terms of mSBD operators:

Lemma 1 (Representation of C-factors using mSBD operator). *Let $\mathbf{S}$ denote a C-component in G. Let $\mathbf{W} \subseteq \mathbf{S}$ denote a set of nodes such that $\mathbf{W} = An(\mathbf{W})_{G(\mathbf{S})}$. Let $\mathbf{R} \equiv Pa(\mathbf{S}) \backslash \mathbf{S}$, and $\mathbf{Z} \equiv (\mathbf{S} \backslash \mathbf{W}) \cap An(\mathbf{R}, \mathbf{W})$. Then,*

1. *$Q[\mathbf{W}] = P_{\mathbf{r}}(\mathbf{w})$;*
2. *$\mathbf{Z}$ satisfies the mSBD criterion relative to $(\mathbf{R}, \mathbf{W})$; and therefore $P_{\mathbf{r}}(\mathbf{w}) = \mathcal{M}[\mathbf{w} \mid \mathbf{r}; \mathbf{z}]$.*

A special case of Lemma 1 is when $\mathbf{W} = \mathbf{S}_i$ for $\mathbf{S}_i$ being a C-component in G, we have $Q[\mathbf{S}_i] = \mathcal{M}[\mathbf{s}_i \mid Pa(\mathbf{s}_i) \cap (\mathbf{v} \backslash \mathbf{s}_i); \emptyset]$. We then propose an identification algorithm that expresses any causal effect as an arithmetic combination of mSBD operators, as shown in Algo. 1. We call the new algorithm *DML-ID* since it will allow us to realize estimators that exhibit DML properties.

DML-ID involves the marginalization of mSBD operators, which can be simplified using the following lemma:

Lemma 2 (Marginalization of mSBD operators). *Let $\mathcal{M}[\mathbf{y} \mid \mathbf{x}; \mathbf{z}]$ be an mSBD operator. For $\mathbf{W} = De(\mathbf{W})_{G[\mathbf{Y}]}$, $\sum_{\mathbf{w}} \mathcal{M}[\mathbf{y} \mid \mathbf{x}; \mathbf{z}] = \mathcal{M}[\mathbf{y} \backslash \mathbf{w} \mid \mathbf{x} \cap Pre(\mathbf{y} \backslash \mathbf{w}); \mathbf{z} \cap Pre(\mathbf{y} \backslash \mathbf{w})]$; For $\mathbf{A} = An(\mathbf{A})_{G[\mathbf{Y}]}$, $\sum_{\mathbf{a}} \mathcal{M}[\mathbf{y} \mid \mathbf{x}; \mathbf{z}] = \mathcal{M}[\mathbf{y} \backslash \mathbf{a} \mid \mathbf{x}; \mathbf{z} \cup \mathbf{a}]$.*

Algorithm 1: DML-ID $(\mathbf{x}, \mathbf{y}, G, P)$

Input: $\mathbf{x}, \mathbf{y}, G(\mathbf{V}), P(\mathbf{v})$.
Output: Expression of $P_{\mathbf{x}}(\mathbf{y})$ as arithmetic combination of mSBD operators; Or FAIL.

1 Let $\mathbf{V} \leftarrow An(\mathbf{Y})$; $P(\mathbf{v}) \leftarrow P(An(\mathbf{Y}))$; and $G \leftarrow G(An(\mathbf{Y}))$.
2 Find the C-components of G: $\mathbf{S}_1, \cdots, \mathbf{S}_{k_s}$.
3 Set $Q[\mathbf{S}_i] = \mathcal{M}[\mathbf{s}_i \mid Pa(\mathbf{s}_i) \cap (\mathbf{v} \backslash \mathbf{s}_i); \emptyset]$. // Lemma 1.
4 Let $\mathbf{D} \equiv An(\mathbf{Y})_{G(\mathbf{V} \backslash \mathbf{X})}$.
5 Find the C-component of $G(\mathbf{D})$: $\mathbf{D}_1, \cdots \mathbf{D}_{k_d}$.
6 For each $\mathbf{D}_j \subseteq \mathbf{S}_i$ for some i, set $Q[\mathbf{D}_j] = \text{MCOMPILE}(\mathbf{D}_j, \mathbf{S}_i, Q[\mathbf{S}_i])$.
7 **return** $P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{d} \backslash \mathbf{y}} \prod_{j=1}^{k_d} Q[\mathbf{D}_j]$.

Procedure MCOMPILE$(\mathbf{C}, \mathbf{T}, Q[\mathbf{T}])$

a.1 Let $\mathbf{A} \equiv An(\mathbf{C})_{G(\mathbf{T})} = \{A_1, A_2, \cdots, A_{n_a}\}$ such that $A_1 \prec A_2 \prec \cdots \prec A_{n_a}$ in $G(\mathbf{T})$.
a.2 Let $Q[\mathbf{A}] = \sum_{\mathbf{T} \backslash \mathbf{A}} Q[\mathbf{T}]$. // Apply Lemma 2 if viable
a.3 **If** $\mathbf{A} = \mathbf{C}$, **then return** $Q[\mathbf{A}]$.
a.4 **If** $\mathbf{A} = \mathbf{T}$, **then return** FAIL.
a.5 **else**
a.6 Let $\mathbf{S}$ be the C-component in $G(\mathbf{A})$ such that $\mathbf{C} \subseteq \mathbf{S}$.
a.7 Let $Q[\mathbf{S}] \equiv \prod_{\{i: A_i \in \mathbf{S}\}} \frac{\sum_{\mathbf{A} \geq i+1} Q[\mathbf{A}]}{\sum_{\mathbf{A} \geq i} Q[\mathbf{A}]}$. // Apply Lemma 2 if viable
a.8 **return** MCOMPILE$(\mathbf{C}, \mathbf{S}, Q[\mathbf{S}])$
end

The sub-procedure MCOMPILE in Algo. 1 derives the expression of the C-factor $Q[\mathbf{D}_j]$ for each $\mathbf{D}_j$ defined in line 5 as an arithmetic combination (marginalization/multiplication/division) of a set of mSBD operators $\{\mathcal{M}_\ell^j\}_{\ell=1}^{m_j}$. We will write $Q[\mathbf{D}_j] = \mathcal{A}^j(\{\mathcal{M}_\ell^j\}_{\ell=1}^{m_j})$, where $\mathcal{A}^j()$ *denote an arithmetic combination operator*.

We show that DML-ID and the original complete algorithm are equivalent in terms of the identification power:

Theorem 1 (Soundness and Completeness of DML-ID). *A causal effect $P_{\mathbf{x}}(\mathbf{y})$ is identifiable if and only if* DML-ID$(\mathbf{x}, \mathbf{y}, G, P)$ *(Algo. 1) returns $P_{\mathbf{x}}(\mathbf{y})$ as an arithmetic combination of mSBD operators, in the form given by*

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{d} \backslash \mathbf{y}} \prod_{j=1}^{k_d} \mathcal{A}^j(\{\mathcal{M}_\ell^j\}_{\ell=1}^{m_j}). \quad (4)$$

We note that Algo. 1 runs in $O\left(|\mathbf{V}|^3\right)$ time, where $|\mathbf{V}|$ denotes the number of variables. A detailed complexity complexity analysis is given in Lemma S.1 in Appendix A.

For concreteness, we demonstrate the application of DML-ID using the models in Fig. (1a,1b), where the effects $P_x(y)$, $P_{x_1,x_2}(y)$ are identifiable by the original identification algorithm as given by Eq. (1) and Eq. (2), respectively.

Demonstration 1 (Algo. 1 for $P_x(y)$ in Example 1 (Fig. 1a)). *We start with $\mathbf{S}_1 = \{W, X, Y\}$ and $\mathbf{S}_2 = \{R\}$*

(Line 2). By Lemma 1, $Q[\mathbf{S}_1] = \mathcal{M}[w,x,y \mid r;\emptyset]$ *and* $Q[\mathbf{S}_2] = \mathcal{M}[r \mid w;\emptyset]$ *(Line 3). Let* $\mathbf{D} = \{Y\}$ *(Line 4,5). Run* mCOMPILE$(Y, \mathbf{S}_1, Q[\mathbf{S}_1])$ *to obtain* $Q[Y]$ *(Line 6). In Procedure* mCOMPILE(), *let* $\mathbf{A}_1 = An(Y)_{G(W,X,Y)} = \{X,Y\}$ *(Line a.1), and* $Q[\mathbf{A}_1] = \sum_w \mathcal{M}[w,x,y \mid r;\emptyset] = \mathcal{M}[x,y \mid r;w] \equiv \mathcal{M}_1$ *by applying the marginalization in Lemma 2 (Line a.2). Let* $\mathbf{S}_Y = \{Y\}$ *(Line a.6). Then,* $Q[Y] = \frac{Q[\mathbf{A}_1]}{\sum_y Q[\mathbf{A}_1]}$, *where* $\sum_y Q[\mathbf{A}_1] = \mathcal{M}[x \mid r;w] \equiv \mathcal{M}_2$ *by Lemma 2 (Line a.7). Finally,* mCOMPILE$(Y, Y, Q[Y])$ *returns* $Q[Y]$ *(Line a.8), and we obtain* $P_x(y) = Q[Y] = \frac{\mathcal{M}_1}{\mathcal{M}_2} \equiv \mathcal{A}(\mathcal{M}_1, \mathcal{M}_2)$ *(Line 7).*

Demonstration 2 (Algo. 1 for $P_{x_1,x_2}(y)$ in Example 2 (Fig. 1b)). *We start with* $\mathbf{S}_1 = \{X_1, Z, Y\}$, $\mathbf{S}_2 = \{R\}$, *and* $\mathbf{S}_3 = \{X_2\}$ *(Line 2). By Lemma 1,* $Q[\mathbf{S}_1] = \mathcal{M}[x_1,z,y \mid (x_2,r);\emptyset]$, $Q[\mathbf{S}_2] = \mathcal{M}[r \mid x_1;\emptyset]$ *and* $Q[\mathbf{S}_3] = \mathcal{M}[x_2 \mid (x_1,z);\emptyset]$ *(Line 3). Let* $\mathbf{D} = \{R,Y\}$ *(Line 4). Let* $\mathbf{D}_1 = \{Y\} \subseteq \mathbf{S}_1$ *and* $\mathbf{D}_2 = \{R\} = \mathbf{S}_2$ *(Line 5). Run* mCOMPILE$(Y, \{\mathbf{S}_1\}, Q[\mathbf{S}_1])$ *to obtain* $Q[Y]$ *(Line 6). Let* $\mathbf{A}_1 = An(Y)_{G(X_1,Z,Y)} = \{Y\}$ *(line a.1) and* $Q[\mathbf{A}_1] = \sum_{x_1,z} \mathcal{M}[x_1,z,y \mid (x_2,r);\emptyset] = \mathcal{M}[y \mid (x_2,r);x_1,z]$ *by Lemma 2 (Line a.2). We obtain* $Q[Y] = Q[\mathbf{A}_1] = \mathcal{M}[y \mid (x_2,r);x_1,z] \equiv \mathcal{M}_1 \equiv \mathcal{A}^1(\mathcal{M}_1)$ *(Line a.3). We obtain* $Q[R] = Q[\mathbf{S}_2] = \mathcal{M}[r \mid x_1;\emptyset] \equiv \mathcal{M}_2 \equiv \mathcal{A}^2(\mathcal{M}_2)$ *(Line 6). Finally, we obtain* $P_{x_1,x_2}(y) = \sum_r \mathcal{A}^1(\mathcal{M}_1)\mathcal{A}^2(\mathcal{M}_2)$ *(Line 7).*

The importance of Thm. 1 lies in that it facilitates deriving an IF for any identified $P_{\mathbf{x}}(\mathbf{y})$ estimands by using the IFs of mSBD operators as a building block.

4 Influence Functions for Causal Estimands

Algo. 1 derives any identifiable causal effects $P_{\mathbf{x}}(\mathbf{y})$ as an arithmetic combinations of mSBDs. In this section, we derive an IF for the identified estimand by first deriving an IF for the mSBD operator. The IF will be used for constructing a DML estimator in the next section.

Lemma 3 (Influence Function for mSBD operator). *Let the target functional be* $\psi \equiv \mathcal{M}[\mathbf{y} \mid \mathbf{x};\mathbf{z}]$. *Then:*

1. $\mathcal{V}_{\mathcal{M}} \equiv \mathcal{V}_{\mathcal{M}}(\{\mathbf{X},\mathbf{Z},\mathbf{Y}\};\eta(P))$ *below is an UIF for* ψ*:*

$$\mathcal{V}_{\mathcal{M}} = H_2 + \sum_{i=2}^{n+1} \mathcal{W}_i \left(H_{i+1} - H_i\right), \tag{5}$$

where, $\mathcal{W}_1 = 1$ *and* $\mathcal{W}_i \equiv \prod_{j=1}^{i-1} \frac{I_{x_j}(X_j)}{P(x_j|\mathbf{x}^{(j-1)},\mathbf{y}^{(j-1)},\mathbf{z}^{(j)})}$ *for* $i = 2,\ldots,n+1$*; and* $H_1 = \psi$, $H_{n+2} \equiv I_{\mathbf{y}}(\mathbf{Y})$, *and* $H_i = P_{\mathbf{x}}\left(\mathbf{y}^{\geq i-1}|\mathbf{Z}^{(i-1)},\mathbf{y}^{(i-2)}\right) I_{\mathbf{y}^{(i-2)}}(\mathbf{Y}^{(i-2)})$ *for* $i = 2,\ldots,n+1$*,*

$$P_{\mathbf{x}}\left(\mathbf{y}^{\geq i-1}|\mathbf{y}^{(i-2)},\mathbf{Z}^{(i-1)}\right) = \sum_{\mathbf{Z}^{\geq i}} \prod_{k=i-1}^{n} q_k(\mathbf{y}|\mathbf{x},\mathbf{Z}) \prod_{j=i}^{n} q_j(\mathbf{Z}|\mathbf{x},\mathbf{y})$$

where $q_k(\mathbf{y}|\mathbf{x},\mathbf{Z}) \equiv P(\mathbf{y}_k|\mathbf{x}^{(k)},\mathbf{y}^{(k-1)},\mathbf{Z}^{(k)})$ *and* $q_j(\mathbf{Z}|\mathbf{x},\mathbf{y}) \equiv P(\mathbf{Z}_j|\mathbf{x}^{(j-1)},\mathbf{y}^{(j-1)},\mathbf{Z}^{(j-1)})$.

2. Let $\mu_{\mathcal{M}} \equiv \mathbb{E}_P[\mathcal{V}_{\mathcal{M}}]$. *Then* $\mu_{\mathcal{M}} = \mathcal{M}[\mathbf{y} \mid \mathbf{x};\mathbf{z}]$.

3. $\phi_{\mathcal{M}} \equiv \phi_{\mathcal{M}}(\{\mathbf{X},\mathbf{Z},\mathbf{Y}\};\psi,\eta(P)) = \mathcal{V}_{\mathcal{M}} - \mu_{\mathcal{M}}$ *is an IF for* ψ.

Algorithm 2: COMPONENTUIF$(\mathcal{A}^j, \mathcal{M}_r^j)$

Input: $\mathcal{A}^j(\{\mathcal{M}_\ell^j\}_{\ell=1}^{m_j})$; $\mathcal{M}_r^j$ for $r \in \{1,\cdots,m_j\}$.
Output: $h_{\mathcal{A}^j,\mathcal{M}_r^j}$

1 Run $h_{\mathcal{A}^j,\mathcal{M}_r^j}(\{\mathcal{M}_\ell^j\}_{\ell=1}^{m_j}, \phi_{\mathcal{M}_r^j}) \leftarrow$ FINDH$(\mathcal{A}^j, \mathcal{M}_r^j)$.
2 $h_{\mathcal{A}^j,\mathcal{M}_r^j} \leftarrow h_{\mathcal{A}^j,\mathcal{M}_r^j}(\{\mu_{\mathcal{M}_\ell^j}\}_{\ell=1}^{m_j}, \mathcal{V}_{\mathcal{M}_r^j} - \mu_{\mathcal{M}_r^j})$ by $\mathcal{M}_\ell^j \leftarrow \mu_{\mathcal{M}_\ell^j}$ and $\phi_{\mathcal{M}_r^j} \leftarrow (\mathcal{V}_{\mathcal{M}_r^j} - \mu_{\mathcal{M}_\ell^j})$.
3 **return** $h_{\mathcal{A}^j,\mathcal{M}_r^j}$

Procedure FINDH$(\mathcal{A}(\{\mathcal{M}_\ell\}), \mathcal{M}_r)$

a.1 Let $\mathcal{A}'(\{\mathcal{M}_\ell\})$, $\mathcal{A}''(\{\mathcal{M}_\ell\})$ denote arithmetic combination operators; let C denote a quantity not involving $\mathcal{M}_r$.
a.2 **if** $\mathcal{A} = C$ **then return** 0.
a.3 **if** $\mathcal{A} = \mathcal{M}_r$ **then return** $\phi_{\mathcal{M}_r}$.
a.4 **if** $\mathcal{A} = C\mathcal{A}'$ **then return** $C *$ FINDH$(\mathcal{A}', \mathcal{M}_r)$.
a.5 **if** $\mathcal{A} = \mathcal{A}'\mathcal{A}''$ **then return** FINDH$(\mathcal{A}', \mathcal{M}_r) * \mathcal{A}'' + \mathcal{A}' *$ FINDH$(\mathcal{A}'', \mathcal{M}_r)$.
a.6 **if** $\mathcal{A} = 1/\mathcal{A}'$ **then return** $-1/(\mathcal{A}')^2 *$ FINDH$(\mathcal{A}', \mathcal{M}_r)$
a.7 **if** $\mathcal{A} = \sum \mathcal{A}'$ **then return** $\sum$FINDH$(\mathcal{A}', \mathcal{M}_r)$.

To derive and represent the IF for the $P_{\mathbf{x}}(\mathbf{y})$ estimand identified by Algo. 1 as given by Eq. (4), we present a couple of useful lemmas next. The first says among the mSBD operators comprising $\mathcal{A}^j(\{\mathcal{M}_\ell^j\}_{\ell=1}^{m_j})$, there exists a special one, named the '*primary mSBD operator* of $\mathcal{A}^j$', as defined in the following:

Lemma 4 (Existence of primary mSBD operator). *Let* $\mathbf{D} = An(\mathbf{Y})_{G(\mathbf{V}\setminus\mathbf{X})}$. *Let* C*-components of* G *be* $\mathbf{S}_i$ *for* $i = 1,2,\cdots,k_s$. *Let* C*-components of* $G(\mathbf{D})$ *be* $\mathbf{D}_j$ *for* $j = 1,2,\cdots,k_d$. *For each* $\mathbf{D}_j \subseteq \mathbf{S}_i$, *let* $Q[\mathbf{D}_j] =$ mCOMPILE$(\mathbf{D}_j, \mathbf{S}_i, Q[\mathbf{S}_i]) = \mathcal{A}^j(\{\mathcal{M}_\ell^j\}_{\ell=1}^{m_j})$. *Then, there exists a* primary mSBD operator, *indexed as* $\mathcal{M}_1^j$ *without loss of generality, such that* $\mathcal{M}_1^j = \mathcal{M}[\mathbf{a}_j \mid Pa(\mathbf{s}_i)\setminus\mathbf{s}_i; \mathbf{s}_i\setminus\mathbf{a}_j]$, *where* $\mathbf{A}_j \equiv An(\mathbf{D}_j)_{G(\mathbf{S}_i)}$.

The following lemma provides an IF of the operator $\mathcal{A}^j$:

Lemma 5 (Influence Function for $Q[\mathbf{D}_j]$). *Let the target functional be* $\psi = Q[\mathbf{D}_j] = \mathcal{A}^j(\{\mathcal{M}_\ell^j\}_{\ell=1}^{m_j})$. *Then, an IF of* ψ *is given by* $\phi_{Q[\mathbf{D}_j]} = \sum_{r=1}^{m_j} h_{\mathcal{A}^j,\mathcal{M}_r^j}$, *where* $h_{\mathcal{A}^j,\mathcal{M}_r^j} =$ COMPONENTUIF$(\mathcal{A}^j, \mathcal{M}_r^j)$ *in Algo. 2.*

We note that Algo. 2 runs in $O\left(m_j^2\right)$ time, where m_j is the number of mSBD operators composing $\mathcal{A}^j$. A detailed analysis is given in Lemma S.2 in Appendix A. The following result gives a special case of Algo. 2.

Corollary 1. *If there are no marginalization operators* $\sum$ *in* $\mathcal{A}^j(\cdot)$, *then* $h_{\mathcal{A}^j,\mathcal{M}_\ell^j} = (\mathcal{V}_{\mathcal{M}_\ell^j} - \mu_{\mathcal{M}_\ell^j})(\partial\mathcal{A}^j(\{\mu_{\mathcal{M}_\ell^j}\}_{\ell=1}^{m_j})/\partial\mu_{\mathcal{M}_\ell^j})$.

We demonstrate Algo. 2 with an example. Assume $\mathcal{A}(\mathcal{M}_1,\mathcal{M}_2) = \mathcal{M}_1/\mathcal{M}_2$, and we derive $h_{\mathcal{A},\mathcal{M}_2}$ by calling COMPONENTUIF$(\mathcal{A},\mathcal{M}_2)$. First FINDH$(\mathcal{A},\mathcal{M}_2)$ is called (line 1). Since $\mathcal{A} = C/\mathcal{M}_2$ for $C = \mathcal{M}_1$, $h_{\mathcal{A},\mathcal{M}_2} =$

$C \cdot$ FINDH$(1/\mathcal{M}_2, \mathcal{M}_2)$ (line a.4). Then, $h_{\mathcal{A},\mathcal{M}_2} = -\mathcal{M}_1/(\mathcal{M}_2)^2 \cdot$ FINDH$(\mathcal{M}_2, \mathcal{M}_2)$ (line a.6), and $h_{\mathcal{A},\mathcal{M}_2} = -\mathcal{M}_1/(\mathcal{M}_2)^2 \cdot \phi_{\mathcal{M}_2}$, where $\phi_{\mathcal{M}_2}$ is IF of $\mathcal{M}_2$ (line a.3). Finally, we obtain $h_{\mathcal{A},\mathcal{M}_2} = -(\mu_{\mathcal{M}_1}/\mu^2_{\mathcal{M}_2})(\mathcal{V}_{\mathcal{M}_2} - \mu_{\mathcal{M}_2})$ (line 2), which is consistent with Coro. 1.

Equipped with Lemmas 4 and 5, an IF for any identifiable causal effects $P_{\mathbf{x}}(\mathbf{y})$ is given as follows:

Theorem 2 (Influence functions for identifiable causal effects). *Let the target functional $\psi \equiv P_{\mathbf{x}}(\mathbf{y})$ be given by Eq. (4). Then, an IF of ψ is given by $\phi_{P_{\mathbf{x}}(\mathbf{y})} = -\psi + \mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}$, where $\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})} \equiv \mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V}; \eta(P))$ is an UIF given by*

$$\begin{aligned}\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})} = &\sum_{\mathbf{d}\setminus\mathbf{y}} \mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1^1}, \{\mu_{\mathcal{M}_l^1}\}_{\ell=2}^{m_1}) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_p}) \\ &+ \sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{\ell=2}^{m_1} h_{\mathcal{A}^1,\mathcal{M}_\ell^1} \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_p}) \\ &+ \sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j=2}^{k_d} \left(\sum_{\ell=1}^{m_j} h_{\mathcal{A}^j,\mathcal{M}_\ell^j} \right) \prod_{\substack{p=1 \\ p\neq j}}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_p}), \quad (6)\end{aligned}$$

where $\mathcal{A}^p(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_p})$ stands for $\mathcal{A}^p(\{\mathcal{M}_\ell^p\}_{\ell=1}^{m_p})$ with $\mathcal{M}_\ell^p$ substituted by $\mu_{\mathcal{M}_\ell^p}$, $\mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1^1}, \{\mu_{\mathcal{M}_l^1}\}_{\ell=2}^{m_1})$ replaces $\mu_{\mathcal{M}_1^1}$ with $\mathcal{V}_{\mathcal{M}_1^1}$, and $h_{\mathcal{A}^j,\mathcal{M}_\ell^j} =$ COMPONENTUIF$(\mathcal{A}^j, \mathcal{M}_\ell^j)$.

We note that Eq. (6) could be derived in $O(|\mathbf{V}|^3)$ time. A detailed complexity analysis is given in Lemma S.3 in Appendix A.

Note in Thm. 2, all $\mathcal{M}_\ell^j$ are replaced with the corresponding $\mu_{\mathcal{M}_\ell^j}$, which is a condition necessary for double robustness. For concreteness, consider the following examples.

Demonstration 3 (Thm. 2 for Example 1). *By Demo. 1, $P_x(y) = Q[Y] = \mathcal{A}(\mathcal{M}_1, \mathcal{M}_2) = \frac{\mathcal{M}_1}{\mathcal{M}_2}$, where $\mathcal{M}_1 = \mathcal{M}[x, y \mid r; w]$ and $\mathcal{M}_2 = \mathcal{M}[x \mid r; w]$. Since $\mathbf{A}_1 = An(Y)_{G(\mathbf{S}_1)} = \{X, Y\}$, $\mathcal{M}_1$ is the primary mSBD operator of $\mathcal{A}$ by Lemma 4. We have $\mathcal{V}_{P_x(y)} = \mathcal{A}(\mathcal{V}_{\mathcal{M}_1}, \mu_{\mathcal{M}_2}) + h_{\mathcal{A},\mathcal{M}_2}$ by Eq. (6), where $\mathcal{A}(\mathcal{V}_{\mathcal{M}_1}, \mu_{\mathcal{M}_2}) = \frac{\mathcal{V}_{\mathcal{M}_1}}{\mu_{\mathcal{M}_2}}$, and $h_{\mathcal{A},\mathcal{M}_2} = -(\mu_{\mathcal{M}_1}/\mu^2_{\mathcal{M}_2})(\mathcal{V}_{\mathcal{M}_2} - \mu_{\mathcal{M}_2})$ by Coro. 1. , or by calling* COMPONENTUIF$(\mathcal{A}, \mathcal{M}_2)$. *Finally, $\phi_{P_x(y)} = -\psi + \mathcal{V}_{P_x(y)}$, where*

$$\mathcal{V}_{P_x(y)} = (1/\mu_{\mathcal{M}_2})\left(\mathcal{V}_{\mathcal{M}_1} - (\mu_{\mathcal{M}_1}/\mu_{\mathcal{M}_2})(\mathcal{V}_{\mathcal{M}_2} - \mu_{\mathcal{M}_2})\right) \quad (7)$$

Demonstration 4 (Thm. 2 for Example 2). *By Demo. 2, $P_{x_1,x_2}(y) = \sum_r \mathcal{A}^1(\mathcal{M}_1)\mathcal{A}^2(\mathcal{M}_2)$ where $\mathcal{A}^1(\mathcal{M}_1) = \mathcal{M}_1 = \mathcal{M}[y \mid (x_2, r); (x_1, z)]$, and $\mathcal{A}^2(\mathcal{M}_2) = \mathcal{M}_2 = \mathcal{M}[r \mid x_1; \emptyset]$. $\mathcal{M}_1$ is the primary mSBD operator of $\mathcal{A}^1$ by Lemma 4 (note $\mathbf{D}_1 = \{Y\}$ and $\mathbf{A}_1 = An(Y)_{\mathbf{S}_1} = \mathbf{D}_1$). We have $\mathcal{V}_{P_{x_1,x_2}(y)} = \sum_r \mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1})\mathcal{A}^2(\mu_{\mathcal{M}_2}) + \sum_r h_{\mathcal{A}^2,\mathcal{M}_2}\mathcal{A}^1(\mu_{\mathcal{M}_1})$ by Eq. (6), where $\mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1}) = \mathcal{V}_{\mathcal{M}_1}$, $\mathcal{A}^2(\mu_{\mathcal{M}_2}) = \mu_{\mathcal{M}_2}$, $\mathcal{A}^1(\mu_{\mathcal{M}_1}) = \mu_{\mathcal{M}_1}$, and $h_{\mathcal{A}^2,\mathcal{M}_2} = \mathcal{V}_{\mathcal{M}_2} - \mu_{\mathcal{M}_2}$ by Coro. 1, or by calling* COMPONENTUIF$(\mathcal{A}^2, \mathcal{M}_2)$. *Finally, $\phi_{P_{x_1,x_2}(y)} = -\psi + \mathcal{V}_{P_{x_1,x_2}(y)}$, where*

$$\mathcal{V}_{P_{x_1,x_2}(y)} = \sum_r (\mathcal{V}_{\mathcal{M}_1}\mu_{\mathcal{M}_2} + (\mathcal{V}_{\mathcal{M}_2} - \mu_{\mathcal{M}_2})\mu_{\mathcal{M}_1}). \quad (8)$$

5 Double Machine Learning Estimators

In this section, we construct DML estimators for any identifiable causal effects $P_{\mathbf{x}}(\mathbf{y})$ from finite samples $\mathcal{D} = \{\mathbf{V}_{(i)}\}_{i=1}^N$, based on the IF discussed above. The resulting DML estimators have nice properties of debiasedness, as well as doubly robustness, in a sense similar to the doubly robustness of BD/SBD estimators found in the literature (Robins, Rotnitzky, and Zhao 1994; Bang and Robins 2005); i.e., an estimator T_N composed of the nuisances $\eta = (\eta_0, \eta_1)$ is said to be doubly robust if T_N is consistent whenever models for either η_0 or η_1 are correctly specified.

Building on (Chernozhukov et al. 2016, Thm. 1), we show that the IF $\phi_{P_{\mathbf{x}}(\mathbf{y})}$ in Thm. 2 is a Neyman orthogonal score:

Proposition 1. *Let the target functional $\psi \equiv P_{\mathbf{x}}(\mathbf{y})$ be given in Eq. (4). The IF $\phi_{P_{\mathbf{x}}(\mathbf{y})}$ for ψ given in Thm. 2 is a Neyman orthogonal score for ψ.*

A DML estimator for $P_{\mathbf{x}}(\mathbf{y})$, named *DML-ID* (DML estimator for any identifiable causal effects), is constructed according to (Chernozhukov et al. 2018) as follows:

Definition 4 (Double Machine Learning Estimator for identifiable causal effects (DML-ID)). Let $\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V}; \eta(P))$ given in Eq. (6) be the UIF for the target functional $\psi = P_{\mathbf{x}}(\mathbf{y})$. Let $\mathcal{D} = \{\mathbf{V}_{(i)}\}_{i=1}^N$ denote samples drawn from $P(\mathbf{v})$. Then, the DML-ID estimator T_N for $\psi = P_{\mathbf{x}}(\mathbf{y})$ is constructed as follows:

1. Split $\mathcal{D}$ randomly into two halves: $\mathcal{D}_0$ and $\mathcal{D}_1$.
2. For $k \in \{0, 1\}$, use $\mathcal{D}_k$ to construct models for $\eta(\widehat{P}_k)$, the nuisance functions estimated from samples $\mathcal{D}_k$.
3. $T_N \equiv \sum_{k\in\{0,1\}} \left(\frac{2}{N} \sum_{\mathbf{V}_{(i)}\in\mathcal{D}_k} \mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V}_{(i)}, \eta(\widehat{P}_{1-k})) \right)$.

We show that DML-ID estimators attain the two aforementioned properties, the main result of this section:

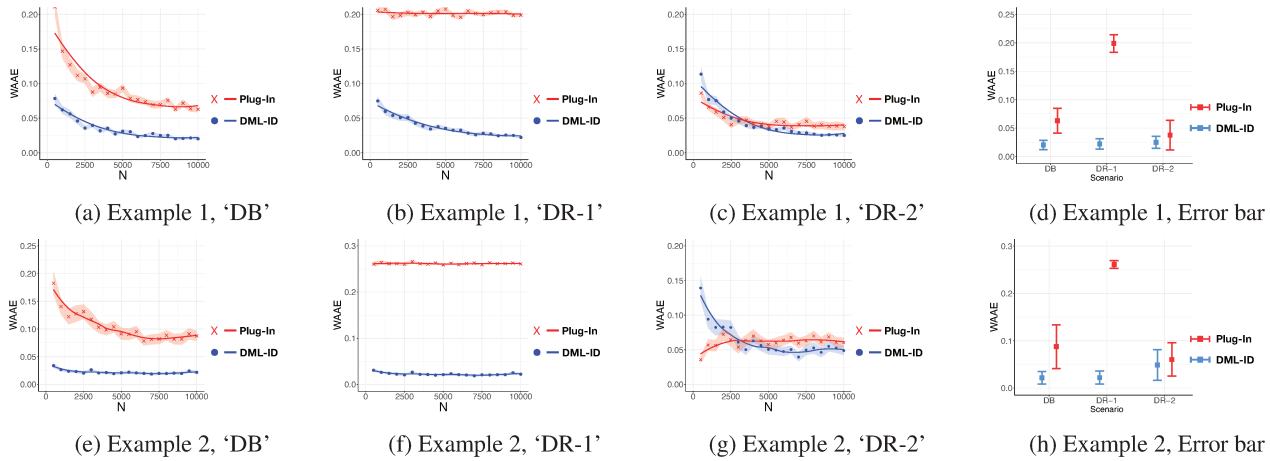
Theorem 3 (Properties of DML-ID). *Let $P_{\mathbf{x}}(\mathbf{y})$ be any identifiable causal effects. Let $\mathbf{D} = An(\mathbf{Y})_{G(\mathbf{V}\setminus\mathbf{X})}$. Let C-components of G be $\mathbf{S}_i$ for $i = 1, 2, \cdots, k_s$. Let C-components of $G(\mathbf{D})$ be $\mathbf{D}_j$ for $j = 1, 2, \cdots, k_d$. For each $\mathbf{D}_j \subseteq \mathbf{S}_i$, let $\mathbf{A}_j \equiv An(\mathbf{D}_j)_{G(\mathbf{S}_i)}$, and let $\mathcal{M}_1^j = \mathcal{M}[\mathbf{y}_j \mid \mathbf{x}_j; \mathbf{z}_j]$ be the primary mSBD operator (defined in Lemma 4), where $\mathbf{X}_j \equiv Pa(\mathbf{S}_i)\setminus\mathbf{S}_i = \{X_{j,i}\}_{i=1}^{n_j}$, $\mathbf{Y}_j \equiv \mathbf{A}_j = \{\mathbf{Y}_{j,i}\}_{i=0}^{n_j}$, $\mathbf{Z}_j \equiv \mathbf{S}_i\setminus\mathbf{A}_j = \{\mathbf{Z}_{j,i}\}_{i=1}^{n_j}$.*

Let T_N be the DML-ID estimator of $P_{\mathbf{x}}(\mathbf{y})$ defined in Def. 4, constructed based on the UIF $\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V}; \eta(P))$. T_N is $\sqrt{N}$-consistent and asymptotically normal if,

1. ***Debiasedness****: Models for all nuisance functions $\eta(P)$ converge at least at rate $o_P(N^{-1/4})$; or*

2. ***Doubly Robustness****: For every $j = 1, 2, \cdots, k_d$, the models for either $\{P(\mathbf{y}_{j,i}|\mathbf{y_j}^{(i-1)}, \mathbf{x_j}^{(i)}, \mathbf{z_j}^{(i)}), P(\mathbf{z}_{j,i}|\mathbf{y_j}^{(i-1)}, \mathbf{x_j}^{(i-1)}, \mathbf{z_j}^{(i-1)})\}_{i=1}^{n^j}$ or $\{P(x_{j,i}|\mathbf{y_j}^{(i-1)}, \mathbf{x_j}^{(i-1)}, \mathbf{z_j}^{(i)})\}_{i=1}^{n^j}$ are correctly specified.*

By virtue of these properties, DML-ID estimators attain root-N consistency even when nuisances converge much slower (say, fourth-root-N) or some nuisances are misspecified, without restricting the complexity of estimation models



(a) Example 1, 'DB' (b) Example 1, 'DR-1' (c) Example 1, 'DR-2' (d) Example 1, Error bar

(e) Example 2, 'DB' (f) Example 2, 'DR-1' (g) Example 2, 'DR-2' (h) Example 2, Error bar

Figure 2: Plots for **(Top)** Example 1, and **(Bottom)** Example 2. **(a,b,c),(e,f,g)** WAAE plots for scenarios 'Debiasedness' ('DB'), 'Doubly Robustness' ('DR-1' and 'DR-2'). **(d,h)** Error bar charts comparing WAAE at $N = 10,000$ for Example (1,2). Shades are representing standard deviation. Plots are best viewed in color.

for nuisances (e.g., Donsker condition). As a result, one can employ flexible ML models (e.g., neural nets) for estimating nuisances in estimating the causal functional.

Demonstration 5 (**Thm. 3 to Example 1**). *The DML-ID estimator T_N for $\psi = P_x(y)$ in Example 1 is constructed based on the UIF in Eq. (7), where $\mathcal{M}_1 = \mathcal{M}[x, y \mid r; w]$ and $\mathcal{M}_2 = \mathcal{M}[x \mid r; w]$. T_N is consistent and asymptotically normal provided that models for nuisance functions $\eta(\widehat{P}) = \{\widehat{P}(y|x,r,w), \widehat{P}(x|r,w), \widehat{P}(r|w), \widehat{P}(w)\}$ converge at least at rate $o_P(N^{-1/4})$. From Demo. 3, the primary mSBD operator is $\mathcal{M}_1$. Then, Thm. 3 states that T_N is consistent if $\{\widehat{P}(x,y|r,w), \widehat{P}(w)\}$ or $\{\widehat{P}(r|w)\}$ are correctly specified; note the correct estimate for $P(x,y|r,w)$ implies the correctness of estimates for $P(x|r,w)$. To compare, we note that a plug-in estimator for Eq. (1) is consistent if $\{\widehat{P}(x,y|r,w), \widehat{P}(w)\}$ are correctly specified.*

Demonstration 6 (**Thm. 3 to Example. 2**). *The DML-ID estimator T_N for $\psi = P_{x_1,x_2}(y)$ is constructed based on the UIF in Eq. (8), where $\mathcal{M}_1 = \mathcal{M}[y \mid (x_2, r); x_1, z]$, $\mathcal{M}_2 = \mathcal{M}[r \mid x_1; \emptyset]$. T_N is consistent and asymptotically normal provided that models for nuisance functions $\eta(\widehat{P}) = \{\widehat{P}(y|r,x_1,x_2,z), \widehat{P}(x_2|z,x_1), \widehat{P}(r|x_1), \widehat{P}(z,x_1)\}$ converge at least at rate $o_P(N^{-1/4})$. Both $\mathcal{M}_1$, $\mathcal{M}_2$ are primary mSBD operators. Thm. 3 states T_N is consistent if $\{\widehat{P}(y|r,x_1,x_2,z), \widehat{P}(z,x_1)\}$ or $\{\widehat{P}(r|x_1), \widehat{P}(x_2|x_1,z)\}$; and $\{\widehat{P}(r|x_1)\}$ or $\{\widehat{P}(x_1)\}$ are correctly specified, which is equivalent to $\{\widehat{P}(y|r,x_1,x_2,z), \widehat{P}(z,x_1)\}$ or $\{\widehat{P}(r|x_1), \widehat{P}(x_2|x_1,z)\}$ be correctly specified. The plug-in estimator for Eq. (2) is consistent if $\{\widehat{P}(y|r,x_1,x_2,z), \widehat{P}(z,x_1), \widehat{P}(r|x_1)\}$ are correctly specified.*

6 Experimental Studies

6.1 Experiments Setup

We evaluate the proposed estimators on the models in Examples 1 and 2. Details of the models and the data-generating process are described in Appendix B. Throughout the experiments, the target causal effect is $\mu(\mathbf{x}) \equiv P_{\mathbf{x}}(Y = 1)$, with ground-truth pre-computed.

We compare DML-ID with **Plug-In Estimator (PI)**, the only viable estimator working for any identifiable causal functional. Nuisance functions are estimated using gradient boosting models called XGBoost (Chen and Guestrin 2016), which is known to be flexible.

Accuracy Measure Given $\mathcal{D}$ with N samples, let $\widehat{\mu}_{\text{DML}}(\mathbf{x})$ and $\widehat{\mu}_{\text{PI}}(\mathbf{x})$ be the estimated $P_{\mathbf{x}}(Y = 1)$ using DML-ID and PI estimators. For each $\widehat{\mu} \in \{\widehat{\mu}_{\text{DML}}(\mathbf{x}), \widehat{\mu}_{\text{PI}}(\mathbf{x})\}$, we assess the quality of the estimator by computing the *weighted average absolute error (WAAE)*, averaged over the density of the intervention $\mathbf{X} = \mathbf{x}$: $\text{WAAE}(\widehat{\mu}) \equiv \sum_{\mathbf{x}} |\widehat{\mu}(\mathbf{x}) - \mu(\mathbf{x})| P_N(\mathbf{x})$, where $P_N(\mathbf{x}) \equiv N_{\mathbf{x}}/N$ for $N_{\mathbf{x}} \equiv \frac{1}{N}\sum_{i=1}^{N} I_{\mathbf{x}}(\mathbf{X}_{(i)})$, following a common practice in statistics in assessing the error of estimates for non-binary treatment (Kennedy et al. 2017; Lee, Kennedy, and Mitra 2020). We run 100 simulations for each $N = \{500, 1000, \cdots, 10000\}$ and take the average of those 100 results. We call plot of the average WAAE vs. the sample size N the WAAE plot.

Simulation Strategy To show debiasedness ('DB') property, we add a 'converging noise' ϵ, decaying at a $N^{-\alpha}$ rate (i.e., $\epsilon \sim \text{Normal}(N^{-\alpha}, N^{-2\alpha})$) for $\alpha = 1/4$, to the estimated nuisance values to control the convergence rate of the estimator for nuisances, following the technique in (Kennedy 2020). We simulate a misspecified model for nuisance functions of the form $P(v_i|\cdot)$ by replacing samples for V_i with randomly generated samples V_i', training the model $\widehat{P}(v_i'|\cdot)$, and using this misspecified nuisance in computing the target functional, following (Kang, Schafer et al. 2007).

6.2 Experimental Results

Debiasedness (DB) The WAAE plots for the debiasedness experiments are shown in Fig. 2 (a) and (e) for Examples 1 and 2, respectively. The DML-ID estimator shows the debiasedness property against the converging noise decaying at $N^{-1/4}$ rates, while the PI estimator converges much slower, for both Examples 1 and 2.

Doubly robustness (DR) The WAAE plots for the doubly robustness experiments are shown in Fig. 2 (b, c) for Example 1 and (f, g) for Examples 2. Two misspecification scenarios are simulated for each example. For Example 1, nuisance $\{P(x, y|r, w), P(w)\}$ are misspecified in 'DR-1', and $\{P(r|w)\}$ is misspecified in 'DR-2'. We note that PI estimator under DR-2 scenario does not have model misspecification since $P(r|w)$ is not a nuisance of PI estimator. For Example 2, nuisance $\{P(y|x_1, x_2, r, z), P(x_1, z)\}$ are misspecified in 'DR-1', and $\{P(r|x_1), P(x_2|x_1, z)\}$ are misspecified in 'DR-2'. The results support the doubly robustness of DML-ID, whereas PI may fail to converge, more prominently, when misspecification is present (i.e., DR-1).

Finally, to further assess the performance of DML-ID when compared against PI, we present the error bar chart of averages and ± 1 standard deviations of WAAEs with the fixed $N = 10,000$ for each of the three scenarios (DB, DR-1, DR-2) in Fig. 2 (d) for Example 1 and in Fig. 2 (h) for Example 2.

We emphasize that the main reason for choosing the plug-in estimator as the baseline for comparison is because it is the only counterpart to DML-ID as an estimator of arbitrary identifiable causal effects. The estimator ('CWO') in (Jung, Tian, and Bareinboim 2020a) covers some special settings and is applicable to Example 1, but not to Example 2. A comparison with CWO on Example 1 is provided in Appendix B.3, showing CWO does not enjoy debiasedness or doubly robustness. Finally, we note that if covariate adjustment is the only way of identifying the causal effect, then DML-ID will reduce to the existing DML estimator. If there are other possible expressions for the causal effect in addition to the covariate adjustment (e.g., front-door), Algo. 1 may output an estimand that is not in the form of covariate adjustment, leading to a different estimator. It's an interesting question to investigate the performances of estimators based on different expressions for the same causal effect.

7 Conclusion

We derived influence functions (Thm. 2) and developed a class of DML estimators, named DML-ID (Def. 4), for any causal effects identifiable given a causal graph. These estimators are guaranteed to have the property of debiasedness and doubly robustness (Thm. 3). Our experimental results demonstrate that DML-ID estimators are significantly more robust against model misspecification and slow convergence rate in learning nuisances compared to the only viable estimator working for any identifiable causal estimand (plug-in estimators). We hope the new machinery developed here will allow empirical scientists to derive more reliable and robust causal effect estimates by integrating modern ML methods that are capable of handling complex, high-dimensional data with causal identification theory.

Acknowledgement

We thank Sanghack Lee, and the reviewers for their valuable feedback helping improving the paper. Elias Bareinboim and Yonghan Jung were partially supported by grants from NSF IIS-1704352 and IIS-1750807 (CAREER). 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.; Correa, J. D.; Ibeling, D.; and Icard, T. 2020. On Pearl's Hierarchy and the Foundations of Causal Inference. Technical Report R-60. Causal Artificial Intelligence Laboratory, Columbia University.

Bareinboim, E.; and Pearl, J. 2012. Causal inference by surrogate experiments: z-identifiability. In *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.

Beinlich, I. A.; Suermondt, H. J.; Chavez, R. M.; and Cooper, G. F. 1989. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *AIME 89*, 247–256. Springer.

Benkeser, D.; Carone, M.; Laan, M. V. D.; and Gilbert, P. 2017. Doubly robust nonparametric inference on the average treatment effect. *Biometrika* 104(4): 863–880.

Bhattacharya, R.; Nabi, R.; and Shpitser, I. 2020. Semiparametric Inference For Causal Effects In Graphical Models With Hidden Variables. *arXiv preprint arXiv:2003.12659* .

Casella, G.; and Berger, R. L. 2002. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

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

Chernozhukov, V.; Demirer, M.; Lewis, G.; and Syrgkanis, V. 2019. Semi-parametric efficient policy learning with continuous actions. In *Advances in Neural Information Processing Systems*, 15065–15075.

Chernozhukov, V.; Escanciano, J. C.; Ichimura, H.; Newey, W. K.; and Robins, J. M. 2016. Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033* .

Farbmacher, H.; Huber, M.; Langen, H.; and Spindler, M. 2020. Causal mediation analysis with double machine learning. *arXiv preprint arXiv:2002.12710* .

Foster, D. J.; and Syrgkanis, V. 2019. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036* .

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

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.

Jaber, A.; Zhang, J.; and Bareinboim, E. 2018. Causal Identification under Markov Equivalence. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*.

Johansson, F. D.; Kallus, N.; Shalit, U.; and Sontag, D. 2018. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598* .

Jung, Y.; Tian, J.; and Bareinboim, E. 2020a. Estimating Causal Effects Using Weighting-Based Estimators. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

Jung, Y.; Tian, J.; and Bareinboim, E. 2020b. Learning causal effects via weighted empirical risk minimization. *Advances in Neural Information Processing Systems* 33.

Kallus, N.; and Uehara, M. 2020. Double Reinforcement Learning for Efficient Off-Policy Evaluation in Markov Decision Processes. In *Proceedings of the 38th International Conference on Machine Learning*.

Kang, J. D.; Schafer, J. L.; et al. 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science* 22(4): 523–539.

Kennedy, E. H. 2018. Efficient nonparametric causal inference with missing exposure information. *arXiv preprint arXiv:1802.08952* .

Kennedy, E. H. 2020. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497* .

Kennedy, E. H.; Ma, Z.; McHugh, M. D.; and Small, D. S. 2017. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 79(4): 1229.

Lee, S.; and Bareinboim, E. 2020. Causal Effect Identifiability under Partial-Observability. In *Proceedings of the 37th International Conference on Machine Learning*.

Lee, S.; Correa, J.; and Bareinboim, E. 2020. Generalized Transportability: Synthesis of Experiments from Heterogeneous Domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

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, Y.; Kennedy, E.; and Mitra, N. 2020. Doubly Robust Nonparametric Instrumental Variable Estimators for Survival Outcomes. *arXiv preprint arXiv:2007.12973* .

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.

Pearl, J.; and Mackenzie, D. 2018. *The book of why: the new science of cause and effect*. Basic Books.

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.

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.

Robins, J. M.; Hernan, M. A.; and Brumback, B. 2000. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* 11(5).

Robins, J. M.; Rotnitzky, A.; and Zhao, L. P. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89(427): 846–866.

Rotnitzky, A.; and Smucler, E. 2019. Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models. *arXiv preprint arXiv:1912.00306* .

Rubin, D. B. 1978. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics* 34–58.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, 3076–3085.

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.

Smucler, E.; Sapienza, F.; and Rotnitzky, A. 2020. Efficient adjustment sets in causal graphical models with hidden variables. *arXiv preprint arXiv:2004.10521* .

Syrgkanis, V.; Lei, V.; Oprescu, M.; Hei, M.; Battocchi, K.; and Lewis, G. 2019. Machine learning estimation of heterogeneous treatment effects with instruments. In *Advances in Neural Information Processing Systems*, 15193–15202.

Tian, J.; and Pearl, J. 2002. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence*, 567–573.

Tian, J.; and Pearl, J. 2003. On the identification of causal effects. Technical Report R-290-L.

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

Van der Vaart, A. W. 2000. *Asymptotic statistics*, volume 3. Cambridge university press.

Zadik, I.; Mackey, L.; and Syrgkanis, V. 2018. Orthogonal Machine Learning: Power and Limitations. In *International Conference on Machine Learning*, 5723–5731.