# Estimating Identifiable Causal Effects through Double Machine Learning

**Yonghan Jung**[1], **Jin Tian**[2], **Elias Bareinboim** [3]
[1] Purdue University
[2] Iowa State University
[3] 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; Díaz and van der Laan 2013; Benkeser et al. 2017; Kennedy et al. 2017; Rotnitzky and Smucler 2020; Smucler, Sapienza, and Rotnitzky 2022; Colangelo and Lee 2020), and more recently, for some specific settings (Toth and van der Laan 2016; Rudolph and van der Laan 2017; Fulcher et al. 2019; Kennedy 2020a; 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). DML estimators have been developed and applied in the context of causal functional estimation in various settings (Toth and van der Laan 2016; Rudolph and van der Laan 2017; Zadik, Mackey, and Syrgkanis 2018; Kennedy 2020a; Kennedy, Lorch, and Small 2019; Syrgkanis et al. 2019; Foster and Syrgkanis 2019; Chernozhukov et al. 2019; Kallus and Uehara 2020; Farbmacher et al. 2020; Colangelo and Lee 2020).

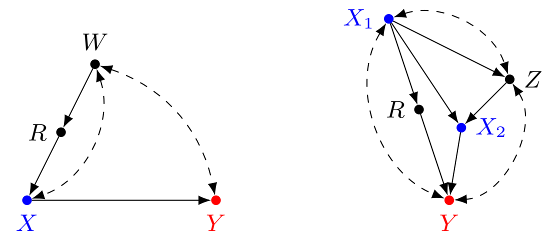
Even though there exists a complete framework for estimating arbitrary identifiable causal functionals based on ML (Jung, Tian, and Bareinboim 2020b), the corresponding procedures do not exhibit DML properties. On the other hand, there are effective and robust estimators for the BD case, which is only 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 $Y$,

[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.

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

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

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 endogenous 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 $Q[\mathbf{C}] \equiv P_{\mathbf{v}\backslash\mathbf{c}}(\mathbf{c})$, called a *$C$-factor*, is defined as the post-intervention distribution of $\mathbf{C}$ under an intervention on $\mathbf{V}\backslash\mathbf{C}$. (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: $P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{d}\backslash\mathbf{y}} Q[\mathbf{D}] = \sum_{\mathbf{d}\backslash\mathbf{y}} \prod_{i=1}^{k_d} Q[\mathbf{D}_i]$, where $\mathbf{D} \equiv An(\mathbf{Y})_{G(\mathbf{V}\backslash\mathbf{X})}$ 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 $t < 1/c$ and $\|g\|_\infty < c$ for some constant $c$ 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 $\phi$), called the *influence function (IF) for the target functional $\psi$*, where $\eta(P)$ stands for the set of nuisance functions comprising $\phi$, 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 $\phi$ 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 $\psi$ (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 2020a).

The treatment provided next assumes that the endogenous variables are discrete, which ascertains that the estimands will be pathwise differentiable. The results can be extended to continuous cases with additional conditions such that the corresponding influence functions are well-defined (Robins 2000; Neugebauer and van der Laan 2007; Díaz and van der Laan 2013; Kennedy et al. 2017; Chernozhukov et al. 2019). We assume the positivity of conditional probabilities as follow: $P(\mathbf{a}|\mathbf{b}) > p_{min} > 0$ for some constant $p_{min} \in (0, 1)$ and for all $\mathbf{a}, \mathbf{b}$ in the support of variables $\mathbf{A}, \mathbf{B} \subseteq \mathbf{V}$.

**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 $\psi$, 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

[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 $\phi$ 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. 2022, 2018).

bias induced by overfitting. DML estimators provide $\sqrt{N}$-consistent estimates of the target $\psi$ 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. 2022).

## 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 for $\psi$ (that turns out to be a Neyman orthogonal score) by first deriving an IF for mSBD estimands and using them as building 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 $\mathbf{R} \equiv Pa(\mathbf{S}) \setminus \mathbf{S}$, and $\mathbf{Z} \equiv (\mathbf{S} \setminus \mathbf{W}) \cap Pre(\mathbf{W})$. 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 $Q[\mathbf{S}_i] = \mathcal{M}[\mathbf{s}_i \mid Pa(\mathbf{s}_i) \cap (\mathbf{v} \setminus \mathbf{s}_i); \emptyset]$. 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 $\mathbf{W} = De(\mathbf{W})_{G[\mathbf{Y}]}$, $\sum_{\mathbf{w}} \mathcal{M}[\mathbf{y} \mid \mathbf{x}; \mathbf{z}] = \mathcal{M}[\mathbf{y} \setminus \mathbf{w} \mid \mathbf{x} \cap Pre(\mathbf{y} \setminus \mathbf{w}); \mathbf{z} \cap Pre(\mathbf{y} \setminus \mathbf{w})]$; For $\mathbf{A} = An(\mathbf{A})_{G[\mathbf{Y}]}$, $\sum_{\mathbf{a}} \mathcal{M}[\mathbf{y} \mid \mathbf{x}; \mathbf{z}] = \mathcal{M}[\mathbf{y} \setminus \mathbf{a} \mid \mathbf{x}; \mathbf{z} \cup \mathbf{a}]$.*

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*

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

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 A.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.* $\mathcal{V}_{\mathcal{M}} \equiv \mathcal{V}_{\mathcal{M}}(\{\mathbf{X}, \mathbf{Z}, \mathbf{Y}\}; \{\pi_0^k, \mu_0^k\}_{k=1}^m)$ *below is an UIF for* $\psi$:

$$\mathcal{V}_{\mathcal{M}} = \overline{\mu}_0^1 + \sum_{k=1}^m \pi_0^{(k)} I_{\mathbf{x}^{(k)}}(\mathbf{X}^{(k)}) \left\{\overline{\mu}_0^{k+1} - \mu_0^k\right\}, \quad (5)$$

*where,* $\overline{\mu}_0^{m+1} \equiv I_{\mathbf{y}}(\mathbf{Y})$, *and for* $k = m, \cdots, 1$,

$$\mu_0^k(\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}) \equiv \mathbb{E}\left[\overline{\mu}_0^{k+1} \middle| \mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}\right],$$

$$\overline{\mu}_0^k(\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)}) \equiv \mathbb{E}\left[\overline{\mu}_0^{k+1} \middle| \mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)}\right].$$

*Also, for* $k = 1, \cdots, m$,

$$\pi_0^k(\mathbf{A}^{(k-1)}, \mathbf{X}^{(k)}) \equiv \frac{1}{P(\mathbf{X}_k | \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)})},$$

$$\pi_0^{(k)}(\mathbf{A}^{(k-1)}, \mathbf{X}^{(k)}) \equiv \prod_{r=1}^k \pi_0^r(\mathbf{A}^{(r-1)}, \mathbf{X}^{(r)}).$$

*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* $\psi$.

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* $\mathbf{A}_j \equiv An(\mathbf{D}_j)_{G(\mathbf{S}_i)}$.

**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 $Q[\mathbf{S}_i] = \mathcal{M}[\mathbf{s}_i \mid Pa(\mathbf{s}_i) \cap (\mathbf{v} \setminus \mathbf{s}_i); \emptyset]$. `// Lemma 1.`
4 Let $\mathbf{D} \equiv An(\mathbf{Y})_{G(\mathbf{V} \setminus \mathbf{X})}$.
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 $Q[\mathbf{D}_j] =$ MCOMPILE$(\mathbf{D}_j, \mathbf{S}_i, Q[\mathbf{S}_i])$.
7 **return** $P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{d} \setminus \mathbf{y}} \prod_{j=1}^{k_d} Q[\mathbf{D}_j]$.

**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 $Q[\mathbf{A}] = \sum_{\mathbf{T} \setminus \mathbf{A}} Q[\mathbf{T}]$. `// 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 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* $\psi$ *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 $O(m_j^2)$ time, where $m_j$ is the number of mSBD operators composing $\mathcal{A}^j$. A detailed analysis is given in Lemma A.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 $\psi$ 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*

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

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

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

Note in Thm. 2, all $\mathcal{M}^j_\ell$ are replaced with the corresponding $\mu_{\mathcal{M}^j_\ell}$, 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 $\text{COMPONENTUIF}(\mathcal{A}, \mathcal{M}_2)$. 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 $\text{COMPONENTUIF}(\mathcal{A}^2, \mathcal{M}_2)$. Finally, $\phi_{P_{x_1,x_2}(y)} = -\psi + \mathcal{V}_{P_{x_1,x_2}(y)}$, where*

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

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

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

1 Run $h_{\mathcal{A}^j,\mathcal{M}^j_r}(\{\mathcal{M}^j_\ell\}_{\ell=1}^{m_j}, \phi_{\mathcal{M}^j_r}) \leftarrow \text{FINDH}(\mathcal{A}^j, \mathcal{M}^j_r)$.
2 $h_{\mathcal{A}^j,\mathcal{M}^j_r} \leftarrow h_{\mathcal{A}^j,\mathcal{M}^j_r}(\{\mu_{\mathcal{M}^j_\ell}\}_{\ell=1}^{m_j}, \mathcal{V}_{\mathcal{M}^j_r} - \mu_{\mathcal{M}^j_r})$ by $\mathcal{M}^j_\ell \leftarrow \mu_{\mathcal{M}^j_\ell}$ and $\phi_{\mathcal{M}^j_r} \leftarrow (\mathcal{V}_{\mathcal{M}^j_r} - \mu_{\mathcal{M}^j_\ell})$.
3 **return** $h_{\mathcal{A}^j,\mathcal{M}^j_r}$
**Procedure** $\text{FINDH}(\mathcal{A}(\{\mathcal{M}_\ell\}), \mathcal{M}_r)$
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** $C * \text{FINDH}(\mathcal{A}', \mathcal{M}_r)$.
a.5 **if** $\mathcal{A} = \mathcal{A}'\mathcal{A}''$ **then return** $\text{FINDH}(\mathcal{A}', \mathcal{M}_r) * \mathcal{A}'' + \mathcal{A}' * \text{FINDH}(\mathcal{A}'', \mathcal{M}_r)$.
a.6 **if** $\mathcal{A} = 1/\mathcal{A}'$ **then return** $-1/(\mathcal{A}')^2 * \text{FINDH}(\mathcal{A}', \mathcal{M}_r)$
a.7 **if** $\mathcal{A} = \sum \mathcal{A}'$ **then return** $\sum \text{FINDH}(\mathcal{A}', \mathcal{M}_r)$.

# 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 robustness properties, which will be exhibited later.

Building on (Chernozhukov et al. 2022, Thm. 1), we show that the IF $\phi_{P_{\mathbf{x}}(\mathbf{y})}$ in Thm. 2 is a Neyman orthogonal score:

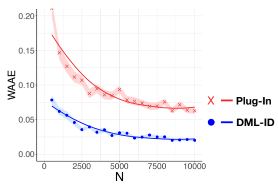
**Proposition 1.** *Let the target functional $\psi \equiv P_{\mathbf{x}}(\mathbf{y})$ be given in Eq. (4). The IF $\phi_{P_{\mathbf{x}}(\mathbf{y})}$ for $\psi$ given in Thm. 2 is a Neyman orthogonal score for $\psi$.*

A DML estimator for $P_{\mathbf{x}}(\mathbf{y})$, named *DML-ID* (DML estimator for any identifiable causal effects), is constructed based on Theorem 2 as follows:

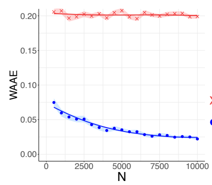
**Definition 4 (DML-ID Estimator).** Let $\mathcal{D} = \{\mathbf{V}_{(i)}\}_{i=1}^N$ denote samples drawn from $P(\mathbf{v})$. Let $\{\mathcal{D}_0, \mathcal{D}_1\}$ denote randomly split two halves of $\mathcal{D}$. Then, the DML-ID (Double Machine Learning estimator for any IDentifiable effect) $T_N$ for $\psi = P_{\mathbf{x}}(\mathbf{y})$ is constructed as follows:

1. For all $j = 1, 2, \cdots, k_d$, $\ell = 1, 2, \cdots, m_j$, estimate $\{\mu_0^{j,\ell,a}, \pi_0^{j,\ell,a}\}_{a=1}^{r_{j,\ell}}$ as $\{\hat{\mu}^{j,\ell,a}, \hat{\pi}^{j,\ell,a}\}_{a=1}^{r_{j,\ell}}$ from $\mathcal{D}_1$ where $\{\mu_0^{j,\ell,a}, \pi_0^{j,\ell,a}\}_{a=1}^{r_{j,\ell}}$ are nuisances of the UIF of mSBD operator $\mathcal{M}^j_\ell$. Evaluate $\hat{\mu}_{\mathcal{M}^j_\ell} \equiv \mathbb{E}_{\mathcal{D}_0}\left[\mathcal{V}_{\mathcal{M}^j_\ell}(\mathbf{V}; \{\hat{\mu}^{j,\ell,a}, \hat{\pi}^{j,\ell,a}\}_{a=1}^{r_{j,\ell}})\right]$ using $\mathcal{D}_0$.
2. Let $T_N(\mathcal{D}_0; \mathcal{D}_1) \equiv \sum_{\mathbf{d}\setminus\mathbf{y}} \prod_{j=1}^{k_d} \mathcal{A}^j(\{\hat{\mu}_{\mathcal{M}^j_\ell}\}_{\ell=1}^{m_j})$.
3. Repeat steps (1-2) after switching $\mathcal{D}_0, \mathcal{D}_1$, and derive $T_N(\mathcal{D}_1; \mathcal{D}_0)$. Then,

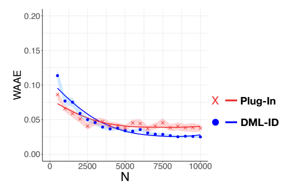
$$T_N = \frac{T_N(\mathcal{D}_0; \mathcal{D}_1) + T_N(\mathcal{D}_1; \mathcal{D}_0)}{2}.$$



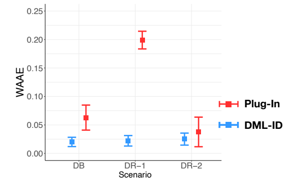

(a) Example 1, ‘DB’



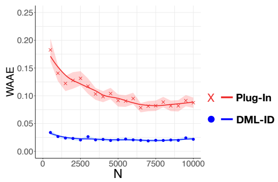

(b) Example 1, ‘DR-1’



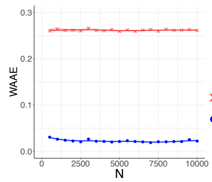

(c) Example 1, ‘DR-2’



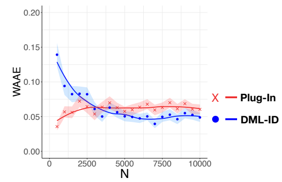

(d) Example 1, Error bar



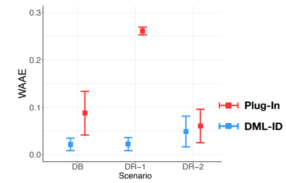

(e) Example 2, ‘DB’




(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.

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 $\{\mathcal{M}^j_\ell\}_{j\in\{1,2,\cdots,k_d\},\ell\in\{1,2,\cdots,m_j\}}$ denote the mSBD adjustments that compose the expression Eq. (4). Let $\{\mu_0^{j,\ell,a}, \pi_0^{j,\ell,a}\}_{a=1}^{r_{j,\ell}}$ denote the set of nuisances constituting the UIF of $\mathcal{M}^j_\ell$ given in Lemma 3, and let $\{\hat{\mu}^{j,\ell,a}, \hat{\pi}^{j,\ell,a}\}_{a=1}^{r_{j,\ell}}$ denote their estimates. Assume that $\hat{\mu}^{j,\ell,a}$ is bounded and $\hat{\pi}^{j,\ell,a}$ is strictly positive and bounded for all $j, \ell, a$. Let $T_N$ be the DML-ID estimator of $P_{\mathbf{x}}(\mathbf{y})$ defined in Def. 4. Then,*

1. ***Debiasedness**:Suppose $\left\|\hat{\mu}^{j,\ell,a} - \mu_0^{j,\ell,a}\right\| = o_P(1)$ and $\left\|\hat{\pi}^{j,\ell,a} - \pi_0^{j,\ell,a}\right\| = o_P(1)$ for all $j, \ell, a$. Then,*

$$T_N - P_{\mathbf{x}}(\mathbf{y}) = R + O_P\left(\sum_{j=1}^{k_d}\sum_{\ell=1}^{m_j}\sum_{a=1}^{r_{j,\ell}}\left\|\hat{\mu}^{j,\ell,a} - \mu_0^{j,\ell,a}\right\|\left\|\hat{\pi}^{j,\ell,a} - \pi_0^{j,\ell,a}\right\|\right), \tag{9}$$

*where $R$ is a variable that converges to a zero-mean normal distribution $\text{NORMAL}(0, \phi^2_{P_{\mathbf{x}}(\mathbf{y})})$ at $\sqrt{N}$ rate, where $\phi_{P_{\mathbf{x}}(\mathbf{y})} = \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V};\eta)$ is the IF of $P_{\mathbf{x}}(\mathbf{y})$ equipped with a true nuisance $\eta$ given in Thm. 2.*

2. ***Doubly Robustness**: If, $\forall j, \ell, a$, either $\hat{\mu}^{j,\ell,a}$ or $\hat{\pi}^{j,\ell,a}$ is correctly specified (i.e., $\hat{\mu}^{j,\ell,a}$ is a consistent estimator for $\mu_0^{j,\ell,a}$ or $\hat{\pi}^{j,\ell,a}$ is a consistent estimator for $\pi_0^{j,\ell,a}$), then $T_N$ is a consistent estimator for $P_{\mathbf{x}}(\mathbf{y})$.*

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 for nuisances (e.g., Donsker condition) (Klaassen 1987; Robins and Ritov 1997; Robins et al. 2008; Zheng and van der Laan 2011; Chernozhukov et al. 2018). 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 using Def. 4. In particular, $P_x(y) = \frac{\mathcal{M}_1}{\mathcal{M}_2}$, where $\mathcal{M}_1 = \mathcal{M}[x, y \mid r; w]$ and $\mathcal{M}_2 = \mathcal{M}[x \mid r; w]$. $\mathcal{V}_{\mathcal{M}_1}$ composes of nuisances $\{\mu_0^1, \pi_0\}$ and $\mathcal{V}_{\mathcal{M}_2}$ composes of $\{\mu_0^2, \pi_0\}$ where $\mu_0^1 \equiv \mathbb{E}[I_{x,y}(X, Y)|R, W] = P(x, y|R, W)$, $\mu_0^2 \equiv \mathbb{E}[I_x(X)|R, W] = P(x|R, W)$, and $\pi_0 \equiv 1/P(R|W)$. Thm. 3 states that $T_N$ converge at $\sqrt{N}$-rate provided that $\hat{\mu}^1$, $\hat{\mu}^2$, $\hat{\pi}$ converge at least at rate $o_P(N^{-1/4})$ to $\mu_0^1, \mu_0^2, \pi_0$. Also, $T_N$ is consistent provided that nuisance estimates $\hat{\mu}^1$ or $\hat{\pi}$; and $\hat{\mu}^2$ or $\hat{\pi}$ are consistent. 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 $\psi = P_{x_1,x_2}(y)$ in Example. 2 is constructed using Def. 4 with $P_{x_1,x_2}(y) = \sum_r \mathcal{M}_1\mathcal{M}_2$, where $\mathcal{M}_1 = \mathcal{M}[y \mid (x_2, r); (x_1, z)]$, $\mathcal{M}_2 = \mathcal{M}[r \mid x_1; \emptyset]$. $\mathcal{V}_{\mathcal{M}_1}$ composes of nuisances $\{\mu_0^{1,1}, \pi_0^{1,1}\}$ and $\mathcal{V}_{\mathcal{M}_2}$ composes of $\{\mu_0^{1,2}, \pi_0^{1,2}\}$ where $\mu_0^{1,1}(R, X_2, Z, X_1) \equiv \mathbb{E}[I_y(Y)|R, X_2, Z, X_1] = P(y|R, X_2, Z, X_1)$, $\pi_0^{1,1}(R, X_2, Z, X_1) = 1/P(R, X_2|Z, X_1)$, $\mu_0^{1,2}(X_1) = \mathbb{E}[I_r(R)|X_1] = P(r|X_1)$, and $\pi_0^{1,2}(X_1) \equiv 1/P(X_1)$. Thm. 3 states that $T_N$ converge at $\sqrt{N}$ rate provided that $\hat{\mu}^{1,1}$, $\hat{\mu}^{1,2}$, $\hat{\pi}^{1,1}$, $\hat{\pi}^{1,2}$ converge at least at rate $o_P(N^{-1/4})$ to $\mu_0^{1,1}, \mu_0^{1,2}, \pi_0^{1,1}, \pi_0^{1,2}$. Also, $T_N$ is consistent provided that nuisance estimates $\hat{\mu}^{1,1}$ or $\hat{\pi}^{1,1}$; and $\hat{\mu}^{1,2}$ or $\hat{\pi}^{1,2}$ are consistent. To compare, we note that a 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 2021). 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' $\epsilon$, 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 2020b). 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 $\{P(r,x_2|x_1,z)\}$ is 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 $\pm 1$ 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 David Childers, Iván Díaz, Uehara Masatoshi, Edward Kennedy, Sanghack Lee, Steve Yadlowsky, and the reviewers for their feedback and help to improve this manuscript. 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. 2022. Locally robust semiparametric estimation. *Econometrica* 90(4): 1501–1535.

Colangelo, K.; and Lee, Y.-Y. 2020. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036* .

Díaz, I.; and van der Laan, M. J. 2013. Targeted data adaptive estimation of the causal dose–response curve. *Journal of Causal Inference* 1(2): 171–192.

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* .

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.

Huang, Y.; and Valtorta, M. 2008. On the completeness of an identifiability algorithm for semi-markovian models. *Annals of Mathematics and Artificial Intelligence* 54(4): 363–408.

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 *Proc. 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. 2020a. Efficient nonparametric causal inference with missing exposure information. *The international journal of biostatistics* 16(1).

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

Kennedy, E. H. 2022. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469* .

Kennedy, E. H.; Balakrishnan, S.; G'Sell, M.; et al. 2020. Sharp instruments for classifying compliers and generalizing causal effects. *Annals of Statistics* 48(4): 2008–2030.

Kennedy, E. H.; Lorch, S.; and Small, D. S. 2019. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81(1): 121–143.

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.

Klaassen, C. A. 1987. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics* 1548–1562.

Koster, J. T.; et al. 2002. Marginalizing and conditioning in graphical models. *Bernoulli* 8(6): 817–840.

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. 2021. Doubly robust nonparametric instrumental variable estimators for survival outcomes. *Biostatistics (Oxford, England)* .

Marsden, J. E.; Hoffman, M. J.; et al. 1993. *Elementary classical analysis*. Macmillan.

Neugebauer, R.; and van der Laan, M. 2007. Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference* 137(2): 419–434.

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.

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.; Li, L.; Tchetgen, E.; van der Vaart, A.; et al. 2008. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, 335–421. Institute of Mathematical Statistics.

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

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

Robins, J. M.; and Ritov, Y. 1997. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Statistics in medicine* 16(3): 285–319.

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. 2020. Efficient Adjustment Sets for Population Average Causal Treatment Effect Estimation in Graphical Models. *Journal of Machine Learning Research* 21(188): 1–86.

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

Rudolph, K. E.; and van der Laan, M. J. 2017. Robust estimation of encouragement-design intervention effects transported across sites. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 79(5): 1509.

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. 2022. Efficient adjustment sets in causal graphical models with hidden variables. *Biometrika* 109(1): 49–65.

Stein, C.; et al. 1956. Efficient nonparametric testing and estimation. In *Proc. of the third Berkeley symposium on mathematical statistics and probability*, volume 1, 187–195.

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.

Toth, B.; and van der Laan, M. 2016. TMLE for marginal structural models based on an instrument. UC Berkeley Division of Biostatistics Working Paper Series. Technical report, working paper 350.

Tsiatis, A. 2007. *Semiparametric theory and missing data*. Springer Science & Business Media.

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.

Zheng, W.; and van der Laan, M. J. 2011. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, 459–474. Springer.

Appendix – Estimating Identifiable Causal Effects through Double Machine Learning

This is a new appendix that includes revised proofs and some new results. Results that only appear in the Appendix will be labeled with 'A' (e.g., Lemma A.1). Otherwise, results will be labeled the same as in the main document.

# A Proofs

## Proof for Time Complexity of Algorithms

**Lemma A.1 (Time complexity of Algo. 1).** *Algo. 1 runs in $O\left(|\mathbf{V}|^3\right)$, where $|\mathbf{V}|$ denote the number of variables.*

*Proof.* In the proof, let $n \equiv |\mathbf{V}|$. Finding $An(\mathbf{Y})$ in line 1 or finding $C$-components in line (2,5) take at most $O(n^2)$, since time complexities for these tasks are bounded by the time for traversing the graph $G$.

Now, we analyze the time complexity of the sub-procedure MCOMPILE for identifying an individual $C$-factor $Q\left[\mathbf{D}_j\right]$ from $Q\left[\mathbf{S}_i\right]$. Let $r_i \equiv |\mathbf{S}_i|$. Then, the number of recursion of MCOMPILE is bounded by $r_i$. For each recursion, the time complexity is $O(n^2)$, for finding $C$-component and ancestral sets. Then, it takes $O(r_i \cdot n^2)$ for identifying an individual $Q\left[\mathbf{D}_j\right]$.

Let $k_d$ be the number of $C$-components, and let $r_1, r_2, \cdots, r_{k_d}$ be sizes of each $C$-components. Then, the time complexity for identifying all $Q\left[\mathbf{D}_1\right], Q\left[\mathbf{D}_2\right], \cdots, Q\left[\mathbf{D}_{k_d}\right]$ is given by

$$O(r_1 \cdot n^2) + O(r_2 \cdot n^2) + \cdots + O(r_{k_d} \cdot n^2) = O\left(n^2 \cdot (r_1 + r_2 + \cdots + r_{k_d})\right) = O(n^3),$$

where the last equality holds since $r_1 + r_2 + \cdots + r_{k_d} = n$. Therefore, we can conclude that Algo. 1 runs in $O(n^3)$, a polynomial time to the size of the graph. □

**Lemma A.2 (Time complexity of Algo. 2).** *Algo. 2 runs in $O\left(m_j^2\right)$.*

*Proof.* Let $m_j$ denote the number of mSBD operators in $\mathcal{A}^j$. Note line 2 of Algo. 2 takes $O(m_j)$. Let the time complexity of the sub-procedure FINDH be $T(m_j)$. Then, the complexity for tasks in line (a.4 - a.7) is given by $T(m_j - 1) + am_j$ for some constant $a$, since those tasks could be done by traversing mSBD operators composing $\mathcal{M}^j$, and invoking the recursion of FINDH whose input size is bounded by $m_j - 1$. Then,

$$T(m_j) = T(m_j - 1) + am_j = T(m_j - 2) + a(m_j - 1) + am_j = \cdots = T(0) + a\left(1 + 2 + \cdots m_j - 1 + m_j\right),$$

where $T(0) = 0$. Since $1 + 2 + \cdots + m_j = \frac{m_j(m_j+1)}{2}$, $T(m_j) = O(m_j^2)$. Therefore, Algo. 2 runs $O\left(m_j^2\right)$. □

**Lemma A.3 (Time complexity for deriving Eq. (6)).** *A closed form of Eq. (6) could be derived in time $O(|\mathbf{V}|^3)$.*

*Proof.* Let $n \equiv |\mathbf{V}|$ in the proof. Running Algo. 1 and obtain Eq. (4) takes $O(n^3)$, as shown in Lemma A.1. By Lemma A.2, it takes $O(m_j^2)$ times to derive $h_{\mathcal{A}^j, m_\ell^j}$ for $\ell \in \{1, 2, \cdots, m_j\}$. This implies that it takes $O(m_j^3)$ time to derive $h_{\mathcal{A}^j, m_\ell^j}$ for all $\ell = 1, 2, \cdots, m_j$. Since the number of $\mathcal{A}^j$ is $k_d$ as in Eq. (4), it takes $O\left(\sum_{j=1}^{k_d} m_j^3\right)$ for deriving all $\{h_{\mathcal{A}^j, \mathcal{M}_\ell^j}\}_{j=1,\cdots,k_d,\ell=1,\cdots,m_j}$.

We now relate $m_j$ with $n$. We note that $m_j$ is bounded by $r_j \equiv |\mathbf{S}_j|$, since line a.7 of Algo. 1 yields at most $|\mathbf{S}|$ number of distinct mSBD operators. Then,

$$O\left(\sum_{j=1}^{k_d} m_j^3\right) = O\left(\sum_{j=1}^{k_d} r_j^3\right) = O\left(\left(\sum_{j=1}^{k_d} r_j\right)^3\right) = O(n^3).$$

Therefore, a time complexity for deriving Eq.(6) is given as $O\left(n^3\right)$. □

## Proof for mSBD Adjustments

**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 $i = 1, \cdots, n$.*

**Proposition A.1 (mSBD adjustment** (Jung, Tian, and Bareinboim 2020a)**).** *If a set of variables $\mathbf{Z} = (\mathbf{Z}_1, \cdots, \mathbf{Z}_m)$ satisfies the mSBD criterion w.r.t. $(\mathbf{X}, \mathbf{Y})$, then the causal effect $P_{\mathbf{x}}(\mathbf{y})$ is given as*

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{z}} \prod_{i=0}^{m} P(\mathbf{y}_i | \mathbf{z}^{(i)}, \mathbf{x}^{(i)}, \mathbf{y}^{(i-1)}) \prod_{j=1}^{m} P(\mathbf{z}_j | \mathbf{z}^{(j-1)}, \mathbf{x}^{(j-1)}, \mathbf{y}^{(j-1)}). \tag{A.10}$$

**Proposition A.2** (**Canonical Expression – Simplified estimand of the mSBD adjustment**). *For the functional in Eq. (A.10), let* $\mathbf{A}_i \equiv \{\mathbf{Y}_i, \mathbf{Z}_{i+1}\}$ *for* $i = 0, \cdots, m$*, where* $\mathbf{Y}_j = \emptyset$ *if* $j < 0$ *and* $\mathbf{Z}_r \equiv \emptyset$ *if* $r \leq 0$*, and* $\mathbf{Z}_{m+1} \equiv \emptyset$*. Let* $\mathbf{A} = \{\mathbf{A}_i\}_{i=0}^m$*. Then, Eq. (A.10) can be represented as*

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{a}'} \prod_{i=0}^{m} P(\mathbf{a}'_i | \mathbf{a}'^{(i-1)}, \mathbf{x}^{(i)}) I_{\mathbf{y}}(\mathbf{y}'), \tag{A.11}$$

*where* $\mathbf{a}'_i \equiv \{\mathbf{y}'_i, \mathbf{z}'_{i+1}\}$ *and* $I_{\mathbf{y}}(\mathbf{y}')$ *is an indicator function, i.e.,* $I_{\mathbf{y}}(\mathbf{y}') = 1$ *if* $\mathbf{y}' = \mathbf{y}$ *and* $0$ *otherwise.*

*Proof.*

$$\begin{aligned}
&\sum_{\mathbf{a}'} \prod_{i=0}^{m} P(\mathbf{a}'_i | \mathbf{a}'^{(i-1)}, \mathbf{x}^{(i)}) I_{\mathbf{y}}(\mathbf{y}') \\
&= \sum_{\mathbf{y}', \mathbf{z}} P(\mathbf{y}'_0, \mathbf{z}_1) P(\mathbf{y}'_1, \mathbf{z}_2 | \mathbf{y}'_0, \mathbf{z}_1, \mathbf{x}_1) \cdots P(\mathbf{y}'_m | \mathbf{y}'^{(m-1)}, \mathbf{z}^{(m)}, \mathbf{x}^{(m)}) I_{\mathbf{y}}(\mathbf{y}') \\
&= \sum_{\mathbf{z}} P(\mathbf{y}_0, \mathbf{z}_1) P(\mathbf{y}_1, \mathbf{z}_2 | \mathbf{y}_0, \mathbf{z}_1, \mathbf{x}_1) \cdots P(\mathbf{y}_m | \mathbf{y}^{(m-1)}, \mathbf{z}^{(m)}, \mathbf{x}^{(m)}) \\
&= \sum_{\mathbf{z}} P(\mathbf{y}_0) P(\mathbf{z}_1 | \mathbf{y}_0) P(\mathbf{y}_1 | \mathbf{y}_0, \mathbf{z}_1, \mathbf{x}_1) P(\mathbf{z}_2 | \mathbf{y}^{(1)}, \mathbf{z}_1, \mathbf{x}_1) \cdots P(\mathbf{y}_m | \mathbf{y}^{(m-1)}, \mathbf{z}^{(m)}, \mathbf{x}^{(m)}) \\
&= \sum_{\mathbf{z}} P(\mathbf{y}_0) P(\mathbf{y}_1 | \mathbf{y}_0, \mathbf{z}_1, \mathbf{x}_1) \cdots P(\mathbf{y}_m | \mathbf{y}^{(m-1)}, \mathbf{z}^{(m)}, \mathbf{x}^{(m)}) \times P(\mathbf{z}_1 | \mathbf{y}_0) \cdots P(\mathbf{z}_m | \mathbf{y}^{(m-1)}, \mathbf{z}^{(m-1)}, \mathbf{x}^{(m-1)}) \\
&= \sum_{\mathbf{z}} \prod_{i=0}^{m} P(\mathbf{y}_i | \mathbf{y}^{(i-1)}, \mathbf{z}^{(i)}, \mathbf{x}^{(i)}) \times \prod_{j=1}^{m} P(\mathbf{z}_j | \mathbf{z}^{(j-1)}, \mathbf{x}^{(j-1)}, \mathbf{y}^{(j-1)}) \\
&= P(\mathbf{y} | do(\mathbf{x})).
\end{aligned}$$

□

## Proof for Lemma 1

**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* $\mathbf{Z} \equiv (\mathbf{S} \backslash \mathbf{W}) \cap Pre(\mathbf{W})$*. 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* $P_{\mathbf{r}}(\mathbf{w}) = \mathcal{M}[\mathbf{w} \mid \mathbf{r}; \mathbf{z}]$.

*Proof.* **First statement**: $P_{\mathbf{v} \backslash \mathbf{w}}(\mathbf{w}) = P_{\mathbf{r}}(\mathbf{w})$.

We first witness $Q[\mathbf{W}] = P_{\mathbf{v} \backslash \mathbf{s}}(\mathbf{w})$. To witness, let $\mathbf{W}' \equiv \mathbf{S} \backslash \mathbf{W}$. Then

$$Q[\mathbf{W}] = P_{\mathbf{v} \backslash \mathbf{w}}(\mathbf{w}) = P_{\mathbf{v} \backslash \mathbf{s}, \mathbf{w}'}(\mathbf{w}) \tag{A.12}$$

$$= P_{\mathbf{v} \backslash \mathbf{s}}(\mathbf{w}). \tag{A.13}$$

Eq. (A.13) follows by applying Rule 3 of do-calculus using the independence $(\mathbf{W} \perp\!\!\!\perp \mathbf{W}' | \mathbf{V} \backslash \mathbf{S})_{G_{\overline{\mathbf{V} \backslash \mathbf{S}, \mathbf{W}'}}}$. We can show that the independence condition holds using contradiction: Assume there exists a path in $G_{\overline{\mathbf{V} \backslash \mathbf{S}, \mathbf{W}'}}$ between $V_i \in \mathbf{W}$ and $V_j \in \mathbf{W}'$. Such path must have arrows going out of $V_j$, the following node in the path must be in $\mathbf{W}$ for the edge in the path to be in $G_{\overline{\mathbf{V} \backslash \mathbf{S}, \mathbf{W}'}}$. But if this is the case, $V_j$ is a parent of some $V_k \in \mathbf{W}$; then $\mathbf{W}$ is not an ancestral set in $G_{\mathbf{S}}$, a contradiction.

Let $\overline{Pa(\mathbf{S})} = Pa(\mathbf{S}) \backslash \mathbf{S}$, which coincides with $\mathbf{R}$. We will use $\overline{Pa(\mathbf{S})}$ and $\mathbf{R}$ interchangeably. We will show $P_{\mathbf{v} \backslash \mathbf{s}}(\mathbf{w}) = P_{\overline{Pa(\mathbf{s})}}(\mathbf{w})$. To show $P_{\mathbf{v} \backslash \mathbf{s}}(\mathbf{w}) = P_{\overline{Pa(\mathbf{s})}}(\mathbf{w})$, we will apply the *do*-calculus Rule 3; $\left(\mathbf{W} \perp\!\!\!\perp \mathbf{V} \backslash Pa(\mathbf{S}) | \overline{Pa(\mathbf{S})}\right)_{G_{\overline{\mathbf{V} \backslash \mathbf{S}}}}$. For any $W_i \in \mathbf{W}$ and $V_j \in \mathbf{V} \backslash Pa(\mathbf{S})$, suppose there is a path between $W_i$ and $V_j$. Since there are no incoming path into $V_j$ in $G_{\overline{\mathbf{V} \backslash \mathbf{S}}}$, the path should have the directed edge from $V_j$ to any node $S_k \in \mathbf{S}$. However, this implies that $V_j \in \overline{Pa(\mathbf{S})}$, which is a contradiction. Notice the path must not be a collider since $\overline{Pa(\mathbf{S})} \subseteq \mathbf{V} \backslash \mathbf{S}$. Therefore, by Rule 3, $P_{\mathbf{v} \backslash \mathbf{s}}(\mathbf{w}) = P_{\overline{Pa(\mathbf{s})}}(\mathbf{w}) = P_{\mathbf{x}}(\mathbf{w})$.

**Second statement**: $\mathbf{Z}$ satisfies the mSBD criterion relative $(\mathbf{R}, \mathbf{W})$.

Let $\mathbf{R} = \{R_1, R_2, \cdots, R_n\}$ where $R_1 \prec R_2 \prec \cdots \prec R_n$. Let $\mathbf{W}_0 \equiv \mathbf{W} \backslash De(\mathbf{R})$, and $\mathbf{W}_i \equiv \mathbf{W} \cap (De(R_i) \backslash De(\mathbf{R}^{\geq i+1}))$ for $i = 1, 2, \cdots, n$. Finally, let $\mathbf{Z}_i \equiv \{V_k \in \mathbf{S} \backslash \mathbf{W} \text{ s.t. } \mathbf{W}_{i-1} \prec V_k \prec R_i\}$ for $i = 1, 2, \cdots, n$. We note that $\mathbf{Z}$ doesn't contain a variable that is a successor of $\mathbf{W}_n$ since $\mathbf{Z}$ is a subset of predecessors of $\mathbf{W}$. Therefore, $\{\mathbf{Z}_1, \cdots, \mathbf{Z}_n\}$ is a partition of $\mathbf{Z}$; i.e., $\mathbf{Z} = \{\mathbf{Z}_1, \cdots, \mathbf{Z}_n\}$.

By such partition, the condition $\mathbf{Z}_i \subseteq ND(\mathbf{R}^{\geq i})$ is automatically satisfied. Thus, we focus on showing

$$\left(\mathbf{W}^{\geq i} \perp\!\!\!\perp R_i | \mathbf{W}^{(i-1)}, \mathbf{Z}^{(i)}, \mathbf{R}^{(i-1)}\right)_{G_{\underline{R_i}\overline{\mathbf{R}^{\geq i+1}}}}. \tag{A.14}$$

Let $G_i \equiv G_{\underline{R_i}\overline{\mathbf{R}^{\geq i+1}}}$. We will show that a path connecting $W_k \in \mathbf{W}^{\geq i}$ and $R_i$ in $G_i$ must be blocked by $\mathbf{W}^{(i-1)}, \mathbf{Z}^{(i)}, \mathbf{R}^{(i-1)}$. To show this, consider a contradictory hypothesis that there is a such path. We note that the path cannot be directed in $G_i$. The path must be either divergent (a path is said to be divergent if it's in a form of $R_i \leftarrow \cdots \leftarrow A \leftrightarrow B \rightarrow \cdots \rightarrow W_k$, where possibly $A = B$), or colliding where the collider is an ancestor of $\mathbf{W}^{(i-1)}, \mathbf{Z}^{(i)}, \mathbf{R}^{(i-1)}$.

Suppose the path is divergent. The path must include a variable in $R_a \in \mathbf{R}$ which has a directed path to a variable in $\mathbf{S}$. Suppose $R_a \in \mathbf{R}^{\geq i+1}$. This means that $R_a$ has a directed path to $R_i$ in $G_i$, which contradicts with the topological order on $\mathbf{R}$. Suppose $R_a \in \mathbf{R}^{(i)}$. Then, if the path is divergent, then the path is blocked by conditioning on $\mathbf{W}^{(i-1)}, \mathbf{Z}^{(i)}, \mathbf{R}^{(i-1)}$.

Suppose the path contains a colliding node $A$ which is an ancestor of $\mathbf{W}^{(i-1)}, \mathbf{Z}^{(i)}, \mathbf{R}^{(i-1)}$. That is, the path contains the subpath s.t. $\rightarrow A \leftarrow \cdots \circ\!\!-\!\!\circ\, W_k$ and $A \rightarrow \cdots \rightarrow V_a$ where $V_a \in \{\mathbf{W}^{(i-1)}, \mathbf{Z}^{(i)}, \mathbf{R}^{(i-1)}\}$. Suppose the subpath connecting $A$ and $W_k$ is directed; i.e., $A \leftarrow \cdots \leftarrow W_k$. Then, $W_k$ becomes an ancestor of $V_a$, which contradicts with the assumed topological order. Therefore, such subpath doesn't exist. Suppose the subpath connecting $A$ and $W_k$ is divergent; i.e., $A \leftarrow \cdots B \leftrightarrow C \rightarrow \cdots \rightarrow W_k$ where $B$ and $C$ are possibly the same node. Such subpath must include a variable $R_a \in \mathbf{R}$. Suppose $R_a \in \mathbf{R}^{\geq i+1}$. This means that $R_a$ has a directed path to $V_a$ in $G_i$, which contradicts with the topological order. Suppose $R_a \in \mathbf{R}^{(i)}$. Then, if the path is divergent, then the path is blocked by conditioning on $\mathbf{W}^{(i-1)}, \mathbf{Z}^{(i)}, \mathbf{R}^{(i-1)}$. Therefore, the subpath is blocked. In conclusion, the path connecting $R_i$ and $W_k$ must be blocked by $\mathbf{W}^{(i-1)}, \mathbf{Z}^{(i)}, \mathbf{R}^{(i-1)}$ in $G_i$. Therefore, Eq. (A.14) holds.

**Main Claim:** If two statements hold, then $Q[\mathbf{W}] = \mathcal{M}[\mathbf{w} \mid \mathbf{r}; \mathbf{z}]$ by the definition of the mSBD adjustment. □

## Proof for Lemma 2

**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 $\mathbf{A} = An(\mathbf{A})_{G[\mathbf{Y}]}$, $\sum_{\mathbf{a}} \mathcal{M}[\mathbf{y} \mid \mathbf{x}; \mathbf{z}] = \mathcal{M}[\mathbf{y}\backslash\mathbf{a} \mid \mathbf{x}; \mathbf{z} \cup \mathbf{a}]$.*

*Proof.* Let $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ and $\mathbf{Y} = \{\mathbf{Y}_k\}_{k=1}^n$. Let $\mathbf{Y}_k = \{Y_{k,1}, \cdots, Y_{k,n_k^y}\}$ where $Y_{k,1} \prec \cdots \prec Y_{k,n_k^y}$; i.e, $Y_{k,a} \prec Y_{k,b}$ if $a < b$ for all $k = 1, 2, \cdots, n$. Then, we represent $\mathbf{Y} = \{\mathbf{Y}_k\}_{k=1}^n = \{\{Y_{k,\ell_k}\}_{\ell_k=1}^{n_k^y}\}_{k=1}^n$ such that $Y_{k_1,\ell_{k_1}} \prec Y_{k_2,\ell_{k_2}}$ whenever $k_1 < k_2$ for any $\ell_{k_1}, \ell_{k_2}$. Then, we can re-index it as $\mathbf{Y} = \{Y_r\}_{r=1}^{n^y}$, where $Y_a \prec Y_b$ whenever $a < b$, by setting $r = (k-1)n_k^y + \ell_k$ for each $\ell_k = 1, 2, \cdots, n_k^y$ for all $k = 1, 2, \cdots, n$.

Let $\mathbf{Z} = \{\mathbf{Z}_p\}_{p=1}^n$. Let $\mathbf{Z}_p = \{Z_{p,1}, \cdots, Z_{p,n_p^z}\}$ where $Z_{p,1} \prec \cdots \prec Z_{p,n_p^z}$; i.e, $Z_{p,a} \prec Z_{p,b}$ if $a < b$ for all $p = 1, 2, \cdots, n$. Then, $\mathbf{Z} = \{\mathbf{Z}_p\}_{p=1}^n = \{\{Z_{p,j_p}\}_{j_p=1}^{n_p^z}\}_{p=1}^n$. Note $Z_{p_1,j_{p_1}} \prec Z_{p_2,j_{p_2}}$ whenever $p_1 < p_2$ for any $j_{p_1}, j_{p_2}$. Then, we can re-index it as $\mathbf{Z} = \{Z_q\}_{q=1}^{n^z}$, where $Z_a \prec Z_b$ whenever $a < b$, by setting $q = (p-1)n_p^z + j_p$ for each $j_p = 1, 2, \cdots, n_p^z$ for all $p = 1, 2, \cdots, n$.

We assume that a topological order for $G$ (denoted $\prec$) is given. For the union $(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z})$ $(= (\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ since they are disjoint, by definition), we consider the $G[\mathbf{X}, \mathbf{Y}, \mathbf{Z}]$. Note the topological order in $G[\mathbf{X}, \mathbf{Y}, \mathbf{Z}]$ could be naturally induced by $\prec$ for $G$.

Let $\mathbf{Y}^{\leq \ell-1}$ denote the set of variables in $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ that are predecessors of $Y_\ell$. $\mathbf{X}^{\leq \ell-1}$ and $\mathbf{Z}^{\leq \ell-1}$ are similarly defined for $X_\ell$ and $Z_\ell$. Also, let $\mathbf{Y}^{\geq \ell}$ denote the set of variables in $\{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$ that are successors of $Y_{\ell-1}$.

For the notational convenience, let $\mathcal{H}_{\mathbf{Y}_k} \equiv \{\mathbf{X}^{(k)}, \mathbf{Y}^{(k-1)}, \mathbf{Z}^{(k-1)}\}$ and $\mathcal{H}_{\mathbf{Z}_k} \equiv \{\mathbf{X}^{(k-1)}, \mathbf{Y}^{(k-1)}, \mathbf{Z}^{(k-1)}\}$.

Note

$$\begin{aligned}
\mathcal{M}[\mathbf{y} \mid \mathbf{x}; \mathbf{z}] &= \sum_{\mathbf{z}} \prod_{\mathbf{y}_k \in \mathbf{Y}} P(\mathbf{y}_k | \mathcal{H}_{\mathbf{y}_k}) \prod_{z_j \in \mathbf{Z}} P(\mathbf{z}_j | \mathcal{H}_{\mathbf{z}_j}) \\
&= \sum_{\mathbf{z}} \prod_{k=0}^{n} \prod_{\ell_k=1}^{n_k^y} P(y_{k,\ell_k} | \mathbf{y}^{\leq (k-1)n_k^y + \ell_k - 1}) \prod_{p=1}^{n} \prod_{j_p=1}^{n_p^z} P(z_{p,j_p} | \mathbf{z}^{\leq (p-1)n_p^z + j_p - 1}) \\
&= \sum_{\mathbf{z}} \prod_{r=1}^{n^y} P(y_r | \mathbf{y}^{\leq r-1}) \prod_{q=1}^{n^z} P(z_q | \mathbf{z}^{\leq q-1}).
\end{aligned}$$

**First statement**: For $\mathbf{W} = De(\mathbf{W})_{G[\mathbf{Y}]}$, $\sum_{\mathbf{w}} \mathcal{M}[\mathbf{y} \mid \mathbf{x}; \mathbf{z}] = \mathcal{M}[\mathbf{y}\backslash\mathbf{w} \mid \mathbf{x} \cap An(\mathbf{w}); \mathbf{z} \cap An(\mathbf{w})]$.

Consider $\mathbf{W} = De(\mathbf{W})_{G[\mathbf{Y}]}$. Since $\mathbf{Y} = \{Y_r\}_{r=1}^{n^y}$ is topologically ordered, we can rewrite it as $\mathbf{W} = \mathbf{Y}^{\geq k_w}$ for some $k_w \leq n_y$. Let $a$ be an index $Z_a \in \mathbf{Z}$ such that $Z_a \prec Y_{k_w-1}$ and $Y_{k_w-1} \prec Z_{a+1}$; i.e., $Z_a$ is the last predecessor of $Y_{k_w}$ in $\mathbf{Z}$. Then,

$$\begin{aligned}\sum_{\mathbf{w}} \mathcal{M}\left[\mathbf{y} \mid \mathbf{x}; \mathbf{z}\right] &= \sum_{\mathbf{w}} \sum_{\mathbf{z}} \prod_{r=1}^{n^y} P(y_r|\mathbf{y}^{\leq r-1}) \prod_{q=1}^{n^z} P(z_q|\mathbf{z}^{\leq q-1}). \\ &= \sum_{\mathbf{z}^{\leq a}} \left( \sum_{\mathbf{y}^{\geq k_w}} \sum_{\mathbf{z}^{\geq a+1}} \prod_{r=1}^{n^y} P(y_r|\mathbf{y}^{\leq r-1}) \prod_{q=1}^{n^z} P(z_q|\mathbf{z}^{\leq q-1}) \right) \\ &= \sum_{\mathbf{z}^{\leq a}} \prod_{r=1}^{k_w-1} P(y_r|\mathbf{y}^{\leq r-1}) \prod_{q=1}^{a} P(z_q|\mathbf{z}^{\leq q-1}). \end{aligned} \tag{A.15}$$

Notice Eq. (A.15) holds, since $(\mathbf{Y}^{\geq k_w}, \mathbf{Z}^{\geq a+1})$ are marginalized out in turn.

Note $\mathbf{Y}^{\leq k_w-1} = \mathbf{Y} \setminus \mathbf{Y}^{\geq k_w} = \mathbf{Y} \setminus \mathbf{W}$. Then, $\mathbf{Z}^{\leq a}$ are the set of predecessors of $\mathbf{Y} \setminus \mathbf{W}$; otherwise, if there exists $Z_q$ for $q \leq a$ such that $Z_q$ is a successor of $\mathbf{Y} \setminus \mathbf{W}$, then such $Z_q$ will be marginalized out. Since Eq. (A.15) only contains conditional probabilities of $(\mathbf{Y} \setminus \mathbf{W})$ and $Pre(\mathbf{Y} \setminus \mathbf{W})$ in $\mathbf{Z}$, none of conditional probabilities in Eq. (A.15) are conditioned on variables in $\mathbf{X} \setminus Pre(\mathbf{Y} \setminus \mathbf{W})$. Therefore,

$$\text{Eq. (A.15)} = \mathcal{M}\left[\mathbf{y} \setminus \mathbf{w} \mid \mathbf{x} \cap Pre(\mathbf{y} \setminus \mathbf{w}); \mathbf{z} \cap Pre(\mathbf{y} \setminus \mathbf{w})\right].$$

**Second statement**: $\mathbf{A} = An(\mathbf{A})_{G[\mathbf{Y}]}$, $\sum_{\mathbf{a}} \mathcal{M}\left[\mathbf{y} \mid \mathbf{x}; \mathbf{z}\right] = \mathcal{M}\left[\mathbf{y} \setminus \mathbf{a} \mid \mathbf{x}; \mathbf{z} \cup \mathbf{a}\right]$.

Consider $\mathbf{A} = An(\mathbf{A})_{G[\mathbf{Y}]}$. Since $\mathbf{Y} = \{Y_r\}_{r=1}^{n^y}$ is topologically ordered, we can rewrite it as $\mathbf{A} = \mathbf{Y}^{\leq k_a}$ for some $k_a \leq n_y$. Let $b$ be the index of $Z_b \in \mathbf{Z}$ and $Y_{k_a} \prec Z_{b+1}$ and $Z_b \prec Y_{k_a}$. Then,

$$\begin{aligned}\sum_{\mathbf{a}} \mathcal{M}\left[\mathbf{y} \mid \mathbf{x}; \mathbf{z}\right] &= \sum_{\mathbf{a}} \sum_{\mathbf{z}} \prod_{r=1}^{n^y} P(y_r|\mathbf{y}^{\leq r-1}) \prod_{q=1}^{n^z} P(z_q|\mathbf{z}^{\leq q-1}) \\ &= \sum_{\mathbf{z}^{\geq b+1}} \sum_{\mathbf{z}^{\leq b}} \sum_{\mathbf{y}^{\leq k_a}} \prod_{r=1}^{n^y} P(y_r|\mathbf{y}^{\leq r-1}) \prod_{q=1}^{n^z} P(z_q|\mathbf{z}^{\leq q-1}). \end{aligned}$$

We note $\sum_{\mathbf{z}^{\leq b}} \sum_{\mathbf{y}^{\leq k_a}}$ does not marginalize out $\mathbf{Z}^{\leq b}$ and $\mathbf{Y}^{\leq K_a}$, since those are predecessors that conditional probabilities $P(y_r|\mathbf{y}^{\leq r-1})$ or $P(z_q|\mathbf{z}^{\leq q-1})$ are dependent on, for some $Y_r$ and $Z_q$. Then,

$$\sum_{\mathbf{a}} \mathcal{M}\left[\mathbf{y} \mid \mathbf{x}; \mathbf{z}\right] = \sum_{\mathbf{z}, \mathbf{y}^{\leq k_a}} \left( \prod_{r=1}^{k_a} \prod_{q=1}^{n^z} P(y_r|\mathbf{y}^{\leq r-1}) P(z_q|\mathbf{z}^{\leq q-1}) \right) \left( \prod_{s=k_a+1}^{n^y} P(y_s|\mathbf{y}^{\leq s-1}) \right) \tag{A.16}$$

Note $\mathbf{Y}^{\leq k_a} = \mathbf{A}$; $(\mathbf{Z}, \mathbf{Y}^{\leq k_a}) = (\{Z_q\}_{q=1}^{n_z}, \{Y_r\}_{r=1}^{k_1}) = \mathbf{Z} \cup \mathbf{A}$; and$\{Y_s\}_{s=k_a+1}^{n^y} = \mathbf{Y} \setminus \mathbf{A}$. Therefore,

$$\sum_{\mathbf{a}} \mathcal{M}\left[\mathbf{y} \mid \mathbf{x}; \mathbf{z}\right] = \mathcal{M}\left[\mathbf{y} \setminus \mathbf{a} \mid \mathbf{x}; \mathbf{z} \cup \mathbf{a}\right].$$

□

## Proof for Theorem 1

**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*

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{d} \setminus \mathbf{y}} \prod_{j=1}^{k_d} \mathcal{A}^j(\{\mathcal{M}_\ell^j\}_{\ell=1}^{m_j}). \tag{A.17}$$

*Proof.* DML-ID follows precisely the original identification algorithm (Alg. 2 in (Tian and Pearl 2003)) except that in Line 3 $Q\left[\mathbf{S}_i\right]$ is expressed in terms of an mSBD operator, which follows from Lemma 1. The soundness and completeness of DML-ID then follows from the soundness and completeness of the original identification algorithm (Huang and Valtorta 2008).

That $\mathbf{D}_j$ is an arithmetic combination (marginalization/multiplication/division) of a set of mSBD operators $\mathcal{A}^j(\{\mathcal{M}_\ell^j\}_{\ell=1}^{m_j})$ is because the only computations invoked by the procedure MCOMPILE() are in Line a.2 (marginalization) and Line a.7 (marginalization, multiplication, and division).

□

**Proof of Lemma 3 (IF & UIF of the mSBD).**

In this paragraph, we provide a proof for deriving an IF & UIF of the mSBD adjustment. The proof of Lemma 3 needs Lemmas (A.4, A.5, A.6, A.7) and Def. A.1.

A parametric submodel is a set of parametric distribuitons $P_\gamma$ s.t. the true distribution belongs to the submodel; i.e., $P = P_{\gamma_0}$ for some $\gamma = \gamma_0$ (Stein et al. 1956). A popular choice of the parametric submodel for the distribution $P$ is

$$P_\gamma(\mathbf{v}) \equiv P(\mathbf{v})\{1 + \gamma g(\mathbf{v})\},$$

where $g(\mathbf{v})$ is a function satisfying $\|g(\mathbf{V})\|_\infty \leq c$ for some constant $c$ so that $P_\gamma(\mathbf{v}) \geq 0$ (Kennedy 2022) and $\mathbb{E}\left[g(\mathbf{V})\right] = 0$.

Let $\nabla_g$ denote the directional derivative along the direction $\gamma$:

$$\nabla_g f(\mathbf{v}) \equiv \frac{\partial}{\partial \gamma} f(\mathbf{v})\{1 + \gamma g(\mathbf{v}))\bigg|_{\gamma=0}.$$

We first derive the Gateaux derivative of conditional distributions.

**Lemma A.4 (Gateaux derivative of conditional distributions).** *Let $\mathbf{V}$ be a set of ordered variables (with an order $\prec$), and $\mathbf{T} \subseteq \mathbf{V}$. For $V_i \subseteq \mathbf{T}$ (i.e., $V_i$ can be a set), the following holds:*

$$\nabla_g P_\gamma(V_i | \mathit{Pre}\,(\mathbf{T})\, V_i) = \left\{\mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T}) | V_i, \mathit{Pre}\,(\mathbf{T})\, V_i\right] - \mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T}) | \mathit{Pre}\,(\mathbf{T})\, V_i\right]\right\} P(V_i | \mathit{Pre}\,(\mathbf{T})\, V_i),$$

*where $S(\mathbf{T}) \equiv \nabla_g \log P_\gamma(\mathbf{T})$.*

*Proof.* Let $S(V_i|\text{pre}_{\mathbf{T}}(V_i)) \equiv \nabla_g \log P_\gamma(V_i|\text{pre}_{\mathbf{T}}(V_i))$. Then,

$$\begin{aligned}
S(V_i|\text{pre}_{\mathbf{T}}(V_i)) &\equiv \nabla_g \log P_\gamma(V_i|\text{pre}_{\mathbf{T}}(V_i)) \\
&= \nabla_g P_\gamma(V_i|\text{pre}_{\mathbf{T}}(V_i)) \underbrace{\frac{\partial}{\partial P(V_i|\text{pre}_{\mathbf{T}}(V_i))} \log P(V_i|\text{pre}_{\mathbf{T}}(V_i))}_{=1/P(V_i|\text{pre}_{\mathbf{T}}(V_i))} \\
&= \nabla_g P_\gamma(V_i|\text{pre}_{\mathbf{T}}(V_i)) \frac{1}{P(V_i|\text{pre}_{\mathbf{T}}(V_i))},
\end{aligned}$$

which implies

$$\nabla_g P_\gamma(V_i|\text{pre}_{\mathbf{T}}(V_i)) = S(V_i|\text{pre}_{\mathbf{T}}(V_i)) P(V_i|\text{pre}_{\mathbf{T}}(V_i)).$$

Therefore, it suffices to show

$$S(V_i|\text{pre}_{\mathbf{T}}(V_i)) = \left\{\mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|V_i, \text{pre}_{\mathbf{T}}(V_i)\right] - \mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|\text{pre}_{\mathbf{T}}(V_i)\right]\right\}.$$

We first note that the mean of the score function is zero, because

$$\begin{aligned}
\mathbb{E}_{P(V_i|\text{pre}_{\mathbf{T}}(V_i))}\left[S(V_i|\text{pre}_{\mathbf{T}}(V_i))\right] &= \sum_{v_i} P(v_i|\text{pre}_{\mathbf{T}}(v_i)) S(v_i|\text{pre}_{\mathbf{T}}(V_i)) \\
&= \sum_{v_i} \cancel{P(v_i|\text{pre}_{\mathbf{T}}(V_i))} \frac{1}{\cancel{P(v_i|\text{pre}_{\mathbf{T}}(V_i))}} \underbrace{\frac{\partial P_\gamma(v_i|\text{pre}_{\mathbf{T}}(V_i))}{\partial \gamma}|_{\gamma=0}}_{=\nabla_g P(v_i|\text{pre}_{\mathbf{T}}(V_i))} \\
&= \frac{\partial}{\partial \gamma} \sum_{v_i} P_\gamma(v_i|\text{pre}_{\mathbf{T}}(V_i))\bigg|_{\gamma=0} \\
&= 0.
\end{aligned}$$

Also, from the fact that $P_\gamma(\mathbf{T}) = \prod_{V_i \in \mathbf{T}} P_\gamma(V_i|\text{pre}_{\mathbf{T}}(V_i))$ (this equality holds since $P_\gamma(\mathbf{T})$ is a valid distribution), we note $S(\mathbf{T}) = \sum_{V_i \in \mathbf{T}} S(V_i|\text{pre}_{\mathbf{T}}(V_i))$. Then, we will study $\mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|V_i, \text{pre}_{\mathbf{T}}(V_i)\right]$ which is decomposed as

$$\begin{aligned}
\mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|V_i, \text{pre}_{\mathbf{T}}(V_i)\right] &= \sum_{V_r \in \mathbf{T}} \mathbb{E}\left[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|V_i, \text{pre}_{\mathbf{T}}(V_i)\right] \\
&= \sum_{\substack{V_r \in \mathbf{T} \\ V_r \succ V_i}} \mathbb{E}\left[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|V_i, \text{pre}_{\mathbf{T}}(V_i)\right] \\
&+ \sum_{\substack{V_r \in \mathbf{T} \\ V_r \prec V_i}} \mathbb{E}\left[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|V_i, \text{pre}_{\mathbf{T}}(V_i)\right] \\
&+ \mathbb{E}\left[S(V_i|\text{pre}_{\mathbf{T}}(V_i))|V_i, \text{pre}_{\mathbf{T}}(V_i)\right].
\end{aligned}$$

For any $V_r \succ V_i$,

$$\begin{aligned}\mathbb{E}_{P(\mathbf{T})}[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|V_i, \text{pre}_{\mathbf{T}}(V_i)] &= \mathbb{E}_{P(\mathbf{T})}[\mathbb{E}_{P(\mathbf{T})}\left[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|\text{pre}_{\mathbf{T}}(V_r)\right]|V_i, \text{pre}_{\mathbf{T}}(V_i)] \\ &= \mathbb{E}_{P(\mathbf{T})}[\underbrace{\mathbb{E}_{P(V_r|\text{pre}_{\mathbf{T}}(V_r))}\left[S(V_r|\text{pre}_{\mathbf{T}}(V_r))\right]}_{=0}|V_i, \text{pre}_{\mathbf{T}}(V_i)] \\ &= 0.\end{aligned}$$

Also,

$$\begin{aligned}\mathbb{E}_{P(\mathbf{T})}[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|\text{pre}_{\mathbf{T}}(V_i)] &= \mathbb{E}_{P(\mathbf{T})}[\mathbb{E}_{P(\mathbf{T})}\left[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|\text{pre}_{\mathbf{T}}(V_r)\right]|\text{pre}_{\mathbf{T}}(V_i)] \\ &= \mathbb{E}_{P(\mathbf{T})}[\underbrace{\mathbb{E}_{P(V_r|\text{pre}_{\mathbf{T}}(V_r))}\left[S(V_r|\text{pre}_{\mathbf{T}}(V_r))\right]}_{=0}|\text{pre}_{\mathbf{T}}(V_i)] \\ &= 0.\end{aligned}$$

For any $V_r \prec V_i$,

$$\begin{aligned}&\mathbb{E}_{P(\mathbf{T})}[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|V_i, \text{pre}_{\mathbf{T}}(V_i)] = S(V_r|\text{pre}_{\mathbf{T}}(V_r)), \text{ and} \\ &\mathbb{E}_{P(\mathbf{T})}[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|\text{pre}_{\mathbf{T}}(V_i)] = S(V_r|\text{pre}_{\mathbf{T}}(V_r)),\end{aligned}$$

since $\{V_r, \text{pre}_{\mathbf{T}}(V_r)\} \subseteq \text{pre}_{\mathbf{T}}(V_i)$. This implies, if $V_r \prec V_i$,

$$\mathbb{E}_{P(\mathbf{T})}[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|V_i, \text{pre}_{\mathbf{T}}(V_i)] = \mathbb{E}_{P(\mathbf{T})}[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|\text{pre}_{\mathbf{T}}(V_i)].$$

Therefore,

$$\begin{aligned}&\mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|V_i, \text{pre}_{\mathbf{T}}(V_i)\right] - \mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|\text{pre}_{\mathbf{T}}(V_i)\right] \\ &= \sum_{V_r \in \mathbf{T}} \left\{\mathbb{E}\left[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|V_i, \text{pre}_{\mathbf{T}}(V_i)\right] - \mathbb{E}\left[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|\text{pre}_{\mathbf{T}}(V_i)\right]\right\} \\ &= \sum_{\substack{V_r \in \mathbf{T} \\ V_r \succ V_i}} \left\{\mathbb{E}\left[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|V_i, \text{pre}_{\mathbf{T}}(V_i)\right] - \mathbb{E}\left[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|\text{pre}_{\mathbf{T}}(V_i)\right]\right\} \\ &+ \sum_{\substack{V_r \in \mathbf{T} \\ V_r \prec V_i}} \left\{\mathbb{E}\left[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|V_i, \text{pre}_{\mathbf{T}}(V_i)\right] - \mathbb{E}\left[S(V_r|\text{pre}_{\mathbf{T}}(V_r))|\text{pre}_{\mathbf{T}}(V_i)\right]\right\} \\ &+ \left\{\mathbb{E}\left[S(V_i|\text{pre}_{\mathbf{T}}(V_i))|V_i, \text{pre}_{\mathbf{T}}(V_i)\right] - \mathbb{E}\left[S(V_i|\text{pre}_{\mathbf{T}}(V_i))|\text{pre}_{\mathbf{T}}(V_i)\right]\right\} \\ &= \mathbb{E}\left[S(V_i|\text{pre}_{\mathbf{T}}(V_i))|V_i, \text{pre}_{\mathbf{T}}(V_i)\right] - \underbrace{\mathbb{E}\left[S(V_i|\text{pre}_{\mathbf{T}}(V_i))|\text{pre}_{\mathbf{T}}(V_i)\right]}_{=0} \\ &= S(V_i|\text{pre}_{\mathbf{T}}(V_i)).\end{aligned}$$

Therefore,

$$S(V_i|\text{pre}_{\mathbf{T}}(V_i)) = \mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|V_i, \text{pre}_{\mathbf{T}}(V_i)\right] - \mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|\text{pre}_{\mathbf{T}}(V_i)\right],$$

and this concludes the proof. □

Using the result, we can derive the influence function of the product of conditional distributions:

**Lemma A.5.** *Let $\mathbf{V}$ be a set of ordered variables (with an order $\prec$), and $\mathbf{T} \subseteq \mathbf{V}$. Suppose $\mathbf{T}$ is decomposed into $\mathbf{T} = \mathbf{A} \cup \mathbf{X}$. Let $P_\pi(\mathbf{T})$ denote the distribution*

$$P_\pi(\mathbf{T}) \equiv \prod_{V_i \in \mathbf{A}} P(V_i|pre_{\mathbf{T}}(V_i)) I_{\mathbf{x}}(\mathbf{X}).$$

*Then, an influence function of the functional*

$$\Psi(P) \equiv \sum_{\mathbf{a}} \prod_{V_i \in \mathbf{A}} P(V_i|pre_{\mathbf{T}}(V_i)) f(\mathbf{a}) \tag{A.18}$$

*is*

$$\phi = \sum_{V_j \in \mathbf{A}} \frac{P_\pi(pre_{\mathbf{T}}(V_j))}{P(pre_{\mathbf{T}}(V_j))} \left\{\mathbb{E}_{P_\pi}\left[f(\mathbf{A})|V_j, pre_{\mathbf{T}}(V_j)\right] - \mathbb{E}_{P_\pi}\left[f(\mathbf{A})|pre_{\mathbf{T}}(V_j)\right]\right\}.$$

*Proof.* We will compute

$$
\begin{aligned}
&\nabla_g \Psi(P_\gamma) \\
&= \nabla_g \sum_{\mathbf{a}} \prod_{V_i \in \mathbf{A}} P(v_i|\text{pre}_{\mathbf{T}}(v_i)) f(\mathbf{a}) \\
&= \sum_{V_j \in \mathbf{A}} \sum_{\mathbf{a}} \{\nabla_g P_\gamma(v_j|\text{pre}_{\mathbf{T}}(v_j))\} \prod_{V_i \in \mathbf{A}\setminus V_j} P(v_i|\text{pre}_{\mathbf{T}}(v_i)) f(\mathbf{a}) \\
&= \sum_{V_j \in \mathbf{A}} \sum_{\mathbf{a}} \left\{\mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|v_j, \text{pre}_{\mathbf{T}}(v_j)\right] - \mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|\text{pre}_{\mathbf{T}}(v_j)\right]\right\} P(v_j|\text{pre}_{\mathbf{T}}(v_j)) \prod_{V_i \in \mathbf{A}\setminus V_j} P(v_i|\text{pre}_{\mathbf{T}}(v_i)) f(\mathbf{a}) \\
&= \sum_{V_j \in \mathbf{A}} \sum_{\mathbf{a}} \left\{\mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|v_j, \text{pre}_{\mathbf{T}}(v_j)\right] - \mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|\text{pre}_{\mathbf{T}}(v_j)\right]\right\} \prod_{V_i \in \mathbf{A}} P(v_i|\text{pre}_{\mathbf{T}}(v_i)) f(\mathbf{a}) \\
&= \sum_{V_j \in \mathbf{A}} \sum_{\mathbf{a},\mathbf{x}'} \left\{\mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|v_j, \text{pre}_{\mathbf{T}}(v_j)\right] - \mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|\text{pre}_{\mathbf{T}}(v_j)\right]\right\} \prod_{V_i \in \mathbf{A}} P(v_i|\text{pre}_{\mathbf{T}}(v_i)) I_{\mathbf{x}}(\mathbf{x}') f(\mathbf{a}) \\
&= \sum_{V_j \in \mathbf{A}} \mathbb{E}_{P_\pi}\left[\left\{\mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|V_j, \text{pre}_{\mathbf{T}}(V_j)\right] - \mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|\text{pre}_{\mathbf{T}}(V_j)\right]\right\} f(\mathbf{A})\right].
\end{aligned}
$$

Let $\mathbf{W}_j \in \{\{V_j, \text{pre}_{\mathbf{T}}(V_j)\}, \{\text{pre}_{\mathbf{T}}(V_j)\}\}$ and $h(\mathbf{t}') \equiv f(\mathbf{a}') I_{\mathbf{x}}(\mathbf{x}')$. Then,

$$
\begin{aligned}
&\mathbb{E}_{P_\pi}\left[\mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|\mathbf{W}_j\right] f(\mathbf{A})\right] \\
&= \sum_{\mathbf{t}'} \mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|\mathbf{w}_j'\right] f(\mathbf{a}') P_\pi(\mathbf{t}') I_{\mathbf{x}}(\mathbf{x}') \\
&= \sum_{\mathbf{t}'\setminus\mathbf{w}_j'} \sum_{\mathbf{w}_j'} \mathbb{E}_{P(\mathbf{T})}\left[S(\mathbf{T})|\mathbf{w}_j'\right] h(\mathbf{t}') P_\pi(\mathbf{t}') \\
&= \sum_{\mathbf{t}'\setminus\mathbf{w}_j'} \sum_{\mathbf{w}_j'} \sum_{\mathbf{t}\setminus\mathbf{w}_j} S(\mathbf{t}\setminus\mathbf{w}_j, \mathbf{w}_j') P(\mathbf{t}\setminus\mathbf{w}_j|\mathbf{w}_j') h(\mathbf{t}') P_\pi(\mathbf{t}') \\
&= \sum_{\mathbf{t}'\setminus\mathbf{w}_j'} \sum_{\mathbf{w}_j'} \sum_{\mathbf{t}\setminus\mathbf{w}_j} S(\mathbf{t}\setminus\mathbf{w}_j, \mathbf{w}_j') P(\mathbf{t}\setminus\mathbf{w}_j|\mathbf{w}_j') h(\mathbf{t}'\setminus\mathbf{w}_j', \mathbf{w}_j') P_\pi(\mathbf{t}'\setminus\mathbf{w}_j'|\mathbf{w}_j') P_\pi(\mathbf{w}_j') \\
&= \sum_{\mathbf{w}_j'} \sum_{\mathbf{t}\setminus\mathbf{w}_j} \left\{\sum_{\mathbf{t}'\setminus\mathbf{w}_j'} h(\mathbf{t}'\setminus\mathbf{w}_j', \mathbf{w}_j') P_\pi(\mathbf{t}'\setminus\mathbf{w}_j'|\mathbf{w}_j')\right\} P_\pi(\mathbf{w}_j') P(\mathbf{t}\setminus\mathbf{w}_j|\mathbf{w}_j') S(\mathbf{t}\setminus\mathbf{w}_j, \mathbf{w}_j') \\
&= \sum_{\mathbf{w}_j'} \sum_{\mathbf{t}\setminus\mathbf{w}_j} \mathbb{E}_{P_\pi}\left[h(\mathbf{T})|\mathbf{w}_j'\right] P_\pi(\mathbf{w}_j') P(\mathbf{t}\setminus\mathbf{w}_j|\mathbf{w}_j') S(\mathbf{t}\setminus\mathbf{w}_j, \mathbf{w}_j') \\
&= \sum_{\mathbf{w}_j'} \sum_{\mathbf{t}\setminus\mathbf{w}_j} \mathbb{E}_{P_\pi}\left[h(\mathbf{T})|\mathbf{w}_j'\right] P_\pi(\mathbf{w}_j') \frac{P(\mathbf{t}\setminus\mathbf{w}_j, \mathbf{w}_j')}{P(\mathbf{w}_j')} S(\mathbf{t}\setminus\mathbf{w}_j, \mathbf{w}_j') \\
&= \sum_{\mathbf{w}_j} \sum_{\mathbf{t}\setminus\mathbf{w}_j} \mathbb{E}_{P_\pi}\left[h(\mathbf{T})|\mathbf{w}_j\right] P_\pi(\mathbf{w}_j) \frac{P(\mathbf{t}\setminus\mathbf{w}_j, \mathbf{w}_j)}{P(\mathbf{w}_j)} S(\mathbf{t}\setminus\mathbf{w}_j, \mathbf{w}_j) \\
&= \sum_{\mathbf{t}} \mathbb{E}_{P_\pi}\left[h(\mathbf{T})|\mathbf{w}_j\right] \frac{P_\pi(\mathbf{w}_j)}{P(\mathbf{w}_j)} S(\mathbf{t}) P(\mathbf{t}) \\
&= \mathbb{E}\left[\frac{P_\pi(\mathbf{W}_j)}{P(\mathbf{W}_j)} \mathbb{E}_{P_\pi}\left[h(\mathbf{T})|\mathbf{W}_j\right] S(\mathbf{T})\right] \\
&= \mathbb{E}\left[\frac{P_\pi(\mathbf{W}_j)}{P(\mathbf{W}_j)} \mathbb{E}_{P_\pi}\left[f(\mathbf{A})|\mathbf{W}_j\right] S(\mathbf{T})\right],
\end{aligned}
$$

where the last equality holds since $\mathbb{E}_{P_\pi}[h(\mathbf{T})|\mathbf{W}_j] = \mathbb{E}_{P_\pi}[f(\mathbf{A})|\mathbf{W}_j]$. Therefore,

$$\begin{aligned}\nabla_g \Psi(P_\gamma) &= \sum_{V_j \in \mathbf{A}} \mathbb{E}\left[\left\{\frac{P_\pi(V_j, \text{pre}_{\mathbf{T}}(V_j))}{P(V_j, \text{pre}_{\mathbf{T}}(V_j))}\mathbb{E}_{P_\pi}[f(\mathbf{A})|V_j, \text{pre}_{\mathbf{T}}(V_j)] - \frac{P_\pi(\text{pre}_{\mathbf{T}}(V_j))}{P(\text{pre}_{\mathbf{T}}(V_j))}\mathbb{E}_{P_\pi}[f(\mathbf{A})|\text{pre}_{\mathbf{T}}(V_j)]\right\} S(\mathbf{T})\right] \\ &\overset{*}{=} \sum_{V_j \in \mathbf{A}} \mathbb{E}\left[\frac{P_\pi(\text{pre}_{\mathbf{T}}(V_j))}{P(\text{pre}_{\mathbf{T}}(V_j))}\left\{\mathbb{E}_{P_\pi}[f(\mathbf{A})|V_j, \text{pre}_{\mathbf{T}}(V_j)] - \mathbb{E}_{P_\pi}[f(\mathbf{A})|\text{pre}_{\mathbf{T}}(V_j)]\right\} S(\mathbf{T})\right].\end{aligned}$$

To witness $\overset{*}{=}$, we will prove the following equality:

$$P_\pi(V_j|\text{pre}_{\mathbf{T}}(V_j)) = P(V_j|\text{pre}_{\mathbf{T}}(V_j)) \text{ for } V_j \in \mathbf{A}. \tag{A.19}$$

To witness Eq. (A.19),

$$\begin{aligned}P_\pi(v_j, \text{pre}_{\mathbf{T}}(v_j)) &= \sum_{v_k : V_k \succ V_j} P_\pi(\mathbf{t}) \\ &= \sum_{v_k : V_k \succ V_j} \prod_{V_r \in \mathbf{A}} P(v_r|\text{pre}_{\mathbf{T}}(v_r)) I_{\mathbf{x}}(\mathbf{X}) \\ &= \prod_{V_r \in \mathbf{A} \cap \{V_a : V_a \preceq V_j\}} P(v_r|\text{pre}_{\mathbf{T}}(v_r)) \prod_{X_r \in \mathbf{X} \cap \{V_a : V_a \preceq V_j\}} I_{x_r}(X_r) \\ &= P(v_j|\text{pre}_{\mathbf{T}}(v_j)) \prod_{V_r \in \mathbf{A} \cap \{V_a : V_a \prec V_j\}} P(v_r|\text{pre}_{\mathbf{T}}(v_r)) \prod_{X_r \in \mathbf{X} \cap \{V_a : V_a \prec V_j\}} I_{x_r}(X_r).\end{aligned}$$

Also,

$$\begin{aligned}P_\pi(\text{pre}_{\mathbf{T}}(v_j)) &= \sum_{v_k : V_k \succeq V_j} P_\pi(\mathbf{t}) \\ &= \sum_{v_k : V_k \succeq V_j} \prod_{V_r \in \mathbf{A}} P(v_r|\text{pre}_{\mathbf{T}}(v_r)) I_{\mathbf{x}}(\mathbf{X}) \\ &= \prod_{V_r \in \mathbf{A} \cap \{V_a : V_a \prec V_j\}} P(v_r|\text{pre}_{\mathbf{T}}(v_r)) \prod_{X_r \in \mathbf{X} \cap \{V_a : V_a \prec V_j\}} I_{x_r}(X_r)\end{aligned}$$

Therefore,

$$P_\pi(v_j|\text{pre}_{\mathbf{T}}(v_j)) = \frac{P_\pi(v_j, \text{pre}_{\mathbf{T}}(v_j))}{P_\pi(\text{pre}_{\mathbf{T}}(v_j))} = P(v_j|\text{pre}_{\mathbf{T}}(v_j)).$$

Then,

$$\frac{P_\pi(V_j, \text{pre}_{\mathbf{T}}(V_j))}{P(V_j, \text{pre}_{\mathbf{T}}(V_j))} = \frac{P_\pi(V_j|\text{pre}_{\mathbf{T}}(V_j))}{P(V_j|\text{pre}_{\mathbf{T}}(V_j))}\frac{P_\pi(\text{pre}_{\mathbf{T}}(V_j))}{P(\text{pre}_{\mathbf{T}}(V_j))} = \frac{P_\pi(\text{pre}_{\mathbf{T}}(V_j))}{P(\text{pre}_{\mathbf{T}}(V_j))}.$$

This concludes the proof that the following is an influence function of $\Psi(P)$:

$$\phi = \sum_{V_j \in \mathbf{A}} \frac{P_\pi(\text{pre}_{\mathbf{T}}(V_j))}{P(\text{pre}_{\mathbf{T}}(V_j))}\left\{\mathbb{E}_{P_\pi}[f(\mathbf{A})|V_j, \text{pre}_{\mathbf{T}}(V_j)] - \mathbb{E}_{P_\pi}[f(\mathbf{A})|\text{pre}_{\mathbf{T}}(V_j)]\right\}.$$

□

Before deriving IF & UIF of the mSBD adjustment, we first define nuisances.

**Definition A.1 (Nuisances for mSBD adjustments).** *Let $\overline{\mu}_0^{m+1} \equiv I_{\mathbf{y}}(\mathbf{Y})$, and for $k = m, \cdots, 0$, we first recursively define nuisances $\mu_0^k, \overline{\mu}_0^k$ as follow:*

$$\mu_0^k(\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}) \equiv \mathbb{E}\left[\overline{\mu}_0^{k+1}\middle|\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}\right],$$

$$\overline{\mu}_0^k(\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)}) \equiv \mathbb{E}\left[\overline{\mu}_0^{k+1}\middle|\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)}\right].$$

*Also, for $k = 1, \cdots, m$, we define*

$$\pi_0^k(\mathbf{A}^{(k-1)}, \mathbf{X}^{(k)}) \equiv \frac{1}{P(\mathbf{X}_k|\mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)})},$$
$$\pi_0^{(k)}(\mathbf{A}^{(k-1)}, \mathbf{X}^{(k)}) \equiv \prod_{r=1}^{k} \pi_0^r(\mathbf{A}^{(r-1)}, \mathbf{X}^{(r)}).$$

The mSBD adjustment can be represented using a defined nuisance:

**Lemma A.6.** *Let $\mu_0^k$ be the nuisance defined in Def. A.1.*

$$\mu_0^0 = \sum_{\mathbf{a}'} \prod_{i=0}^{m} P(\mathbf{a}'_i|\mathbf{a}'^{(i-1)}, \mathbf{x}^{(i)}) I_{\mathbf{y}}(\mathbf{y}').$$

*Proof.* We will first prove by induction that the following holds for $k = m, m-1, \cdots, 1$,

$$\overline{\mu}_0^k(\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)}) = \sum_{\mathbf{a}'_k, \mathbf{a}'_{k+1}, \cdots, \mathbf{a}'_m} I_{\mathbf{y}}(\mathbf{y}'^{(k:m)}, \mathbf{Y}^{(k-1)}) \prod_{r=k}^{m} P(\mathbf{a}'_r|\mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)}, \mathbf{x}^{(k:r)}, \mathbf{a}'^{(k:r-1)}), \tag{A.20}$$

where $\mathbf{a}'^{(k:r-1)} = \emptyset$ if $k > r-1$. We first check that Eq. (A.20) holds for the base case $k = m$.

$$\begin{aligned}
\overline{\mu}_0^m(\mathbf{x}_m, \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-1)}) &= \mathbb{E}\left[I_{\mathbf{y}}(\mathbf{Y}) \middle| \mathbf{x}_m, \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-1)}\right] \\
&= \sum_{\mathbf{a}'_m} I_{\mathbf{y}}(\mathbf{y}'_m, \mathbf{Y}^{(m-1)}) P(\mathbf{a}'_m|\mathbf{x}_m, \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-1)}).
\end{aligned}$$

Assume Eq. (A.20) holds for $k$ for the induction step. Then

$$\begin{aligned}
&\overline{\mu}_0^{k-1}(\mathbf{x}_{k-1}, \mathbf{X}^{(k-2)}, \mathbf{A}^{(k-2)}) \\
&\equiv \mathbb{E}\left[\overline{\mu}_0^k(\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)})|\mathbf{x}_{k-1}, \mathbf{X}^{(k-2)}, \mathbf{A}^{(k-2)}\right] \\
&= \mathbb{E}\left[\overline{\mu}_0^k(\mathbf{x}_k, \mathbf{x}_{k-1}, \mathbf{X}^{(k-2)}, \mathbf{A}^{(k-1)})|\mathbf{x}_{k-1}, \mathbf{X}^{(k-2)}, \mathbf{A}^{(k-2)}\right] \\
&= \sum_{\mathbf{a}'_{k-1}} \overline{\mu}_0^k(\mathbf{x}_k, \mathbf{x}_{k-1}, \mathbf{X}^{(k-2)}, \mathbf{a}'_{k-1}, \mathbf{A}^{(k-2)}) P(\mathbf{a}'_{k-1}|\mathbf{x}_{k-1}, \mathbf{X}^{(k-2)}, \mathbf{A}^{(k-2)}) \\
&= \sum_{\mathbf{a}'_{k-1}} \left\{ \sum_{\mathbf{a}'_k, \cdots, \mathbf{a}'_m} I_{\mathbf{y}}(\mathbf{y}'^{(k-1:m)}, \mathbf{Y}^{(k-2)}) \prod_{r=k}^{m} P(\mathbf{a}'_r|\mathbf{X}^{(k-2)}, \mathbf{A}^{(k-2)}, \mathbf{x}^{(k-1:r)}, \mathbf{a}'^{(k-1:r-1)}) \right\} \\
&\times P(\mathbf{a}'_{k-1}|\mathbf{x}_{k-1}, \mathbf{X}^{(k-2)}, \mathbf{A}^{(k-2)}) \\
&= \sum_{\mathbf{a}'_{k-1}, \cdots, \mathbf{a}'_m} I_{\mathbf{y}}(\mathbf{y}'^{(k-1:m)}, \mathbf{Y}^{(k-2)}) \prod_{r=k-1}^{m} P(\mathbf{a}'_r|\mathbf{x}^{(k-1:r)}, \mathbf{X}^{(k-2)}, \mathbf{a}'^{(k-1:r-1)}, \mathbf{A}^{(k-2)}),
\end{aligned}$$

which certifies that Eq. (A.20) holds for every $k$. Choosing $k = 1$ in Eq. (A.20), we have

$$\overline{\mu}_0^1(\mathbf{x}_1, \mathbf{A}_0) = \sum_{\mathbf{a}'_1, \cdots, \mathbf{a}'_m} I_{\mathbf{y}}(\mathbf{y}'^{(1:m)}, \mathbf{Y}_0) \prod_{r=1}^{m} P(\mathbf{a}'_r|\mathbf{x}^{(1:r)}, \mathbf{a}'^{(1:r-1)}, \mathbf{A}_0).$$

By taking an expectation on both sides, we have

$$\mu_0^0 \equiv \mathbb{E}\left[\overline{\mu}_0^1(\mathbf{x}_1, \mathbf{A}_0)\right] = \sum_{\mathbf{a}'} I_{\mathbf{y}}(\mathbf{y}') \prod_{r=0}^{m} P(\mathbf{a}'_r|\mathbf{x}^{(r)}, \mathbf{a}'^{(r-1)})$$

□

**Lemma A.7 (Equivalence between two target quantities).** *The quantity $\Psi(P)$ in Eq. (A.18) in Lemma A.5 can be reduced to Eq. (A.11) by setting $\mathbf{A} = \{\mathbf{A}_i\}_{i=0}^m$ with $\mathbf{A}_i \equiv \{\mathbf{Y}_i, \mathbf{Z}_{i+1}\}$ and $f(\mathbf{a}') = I_{\mathbf{y}}(\mathbf{y}')$, and ordering the variables in $\mathbf{T} = \mathbf{A} \cup \mathbf{X}$ as $\mathbf{A}_0 \prec \mathbf{X}_1 \prec \mathbf{A}_1 \prec \mathbf{X}_2 \prec \cdots \prec \mathbf{X}_m \prec \mathbf{A}_m$. In particular, under such a setting, $pre_{\mathbf{T}}(\mathbf{A}_i) = \{\mathbf{A}^{(i-1)}, \mathbf{X}^{(i)}\}$, and*

$$\Psi(P) \equiv \sum_{\mathbf{a}} \prod_{i:V_i \in \mathbf{A}} P(v_i|pre_{\mathbf{T}}(v_i)) f(\mathbf{a}) = \sum_{\mathbf{a}'} \prod_{i=0}^{m} P(\mathbf{a}'_i|\mathbf{a}'^{(i-1)}, \mathbf{x}^{(i)}) I_{\mathbf{y}}(\mathbf{y}') =: \textit{Eq. (A.11)}.$$

*Proof.* It is clear that, under the order $\mathbf{A}_0 \prec \mathbf{X}_1 \prec \mathbf{A}_1 \prec \mathbf{X}_2 \prec \cdots \prec \mathbf{X}_m \prec \mathbf{A}_m$, We have $\text{pre}_{\mathbf{T}}(\mathbf{A}_i) = \{\mathbf{A}^{(i-1)}, \mathbf{X}^{(i)}\}$. Then, $\Psi(P)$ becomes

$$\begin{aligned}\Psi(P) &\equiv \sum_{\mathbf{a}'} \prod_{\mathbf{A}_i \in \mathbf{A}} P(\mathbf{a}'_i | \text{pre}_{\mathbf{T}}(\mathbf{a}'_i)) I_{\mathbf{y}}(\mathbf{y}') \\ &= \sum_{\mathbf{a}'} \prod_{\mathbf{A}_i \in \mathbf{A}} P(\mathbf{a}'_i | \mathbf{x}^{(i)}, \mathbf{a}'^{(i-1)}) I_{\mathbf{y}}(\mathbf{y}') \\ &= \text{Eq. (A.11)}.\end{aligned}$$

□

**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. $\mathcal{V}_{\mathcal{M}} \equiv \mathcal{V}_{\mathcal{M}}(\{\mathbf{X}, \mathbf{Z}, \mathbf{Y}\}; \{\pi_0^k, \mu_0^k\}_{k=1}^m)$ *below is an UIF for* $\psi$:

$$\mathcal{V}_{\mathcal{M}} = \overline{\mu}_0^1 + \sum_{k=1}^{m} \pi_0^{(k)} I_{\mathbf{x}^{(k)}}(\mathbf{X}^{(k)}) \left\{\overline{\mu}_0^{k+1} - \mu_0^k\right\}, \tag{A.21}$$

*where,* $\overline{\mu}_0^{m+1} \equiv I_{\mathbf{y}}(\mathbf{Y})$, *and for* $k = m, \cdots, 1$,

$$\begin{aligned}\mu_0^k(\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}) &\equiv \mathbb{E}\left[\overline{\mu}_0^{k+1} \middle| \mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}\right], \\ \overline{\mu}_0^k(\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)}) &\equiv \mathbb{E}\left[\overline{\mu}_0^{k+1} \middle| \mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)}\right].\end{aligned}$$

*Also, for* $k = 1, \cdots, m$,

$$\begin{aligned}\pi_0^k(\mathbf{A}^{(k-1)}, \mathbf{X}^{(k)}) &\equiv \frac{1}{P(\mathbf{X}_k | \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)})}, \\ \pi_0^{(k)}(\mathbf{A}^{(k-1)}, \mathbf{X}^{(k)}) &\equiv \prod_{r=1}^{k} \pi_0^r(\mathbf{A}^{(r-1)}, \mathbf{X}^{(r)}).\end{aligned}$$

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* $\psi$.

*Proof.* We will prove the first and the third statements simultaneously. By applying Lemma A.5 and A.7, an IF for the mSBD adjustment in Eq. (A.11) is given by

$$\phi = \sum_{j=0}^{m} \frac{P_\pi(\text{pre}_{\mathbf{T}}(\mathbf{A}_j))}{P(\text{pre}_{\mathbf{T}}(\mathbf{A}_j))} \left\{\mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y}) | \mathbf{A}_j, \text{pre}_{\mathbf{T}}(\mathbf{A}_j)\right] - \mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y}) | \text{pre}_{\mathbf{T}}(\mathbf{A}_j)\right]\right\}$$

where

$$\frac{P_\pi(\text{pre}_{\mathbf{T}}(\mathbf{A}_j))}{P(\text{pre}_{\mathbf{T}}(\mathbf{A}_j))} = \frac{P_\pi(\mathbf{A}^{(j-1)}, \mathbf{X}^{(j)})}{P(\mathbf{A}^{(j-1)}, \mathbf{X}^{(j)})} \stackrel{*}{=} \prod_{k=1}^{j} \pi_0^k(\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}) I_{\mathbf{x}_k}(\mathbf{X}_k) = \pi_0^{(j)}(\mathbf{X}^{(j)}, \mathbf{A}^{(j-1)}) I_{\mathbf{x}^{(j)}}(\mathbf{X}^{(j)}), \tag{A.22}$$

where the equation $\stackrel{*}{=}$ holds since

$$\begin{aligned}\frac{P_\pi(\mathbf{A}^{(j-1)}, \mathbf{X}^{(j)})}{P(\mathbf{A}^{(j-1)}, \mathbf{X}^{(j)})} &= \frac{\sum_{\mathbf{a}^{\geq j}, \mathbf{x}^{\geq j+1}} P_\pi(\mathbf{A}, \mathbf{X})}{\sum_{\mathbf{a}^{\geq j}, \mathbf{x}^{\geq j+1}} P(\mathbf{A}, \mathbf{X})} \\ &= \frac{\sum_{\mathbf{a}^{\geq j}, \mathbf{x}^{\geq j+1}} \prod_{r=0}^{m} P(\mathbf{A}_r | \mathbf{A}^{(r-1)}, \mathbf{X}^{(r)}) \prod_{q=1}^{m} I_{x_q}(X_q)}{\sum_{\mathbf{a}^{\geq j}, \mathbf{x}^{\geq j+1}} \prod_{r=0}^{m} P(\mathbf{A}_r | \mathbf{A}^{(r-1)}, \mathbf{X}^{(r)}) \prod_{q=1}^{m} P(X_q | \mathbf{X}^{(q-1)}, \mathbf{A}^{(q-1)})} \\ &= \frac{\prod_{r=0}^{j-1} P(\mathbf{A}_r | \mathbf{A}^{(r-1)}, \mathbf{X}^{(r)}) \prod_{q=1}^{j} I_{x_q}(X_q)}{\prod_{r=0}^{j-1} P(\mathbf{A}_r | \mathbf{A}^{(r-1)}, \mathbf{X}^{(r)}) \prod_{q=1}^{j} P(X_q | \mathbf{X}^{(q-1)}, \mathbf{A}^{(q-1)})} \\ &= \prod_{q=1}^{j} \frac{I_{x_q}(X_q)}{P(X_q | \mathbf{X}^{(q-1)}, \mathbf{A}^{(q-1)})} \\ &= \prod_{q=1}^{j} \pi_0^q(\mathbf{X}^{(q)}, \mathbf{A}^{(q-1)}) I_{\mathbf{x}_q}(\mathbf{X}_q).\end{aligned}$$

Therefore,

$$\begin{aligned}\phi &= \sum_{j=0}^{m} \frac{P_\pi(\text{pre}_{\mathbf{T}}(\mathbf{A}_j))}{P(\text{pre}_{\mathbf{T}}(\mathbf{A}_j))} \left\{\mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\mathbf{A}_j, \text{pre}_{\mathbf{T}}(\mathbf{A}_j)\right] - \mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\text{pre}_{\mathbf{T}}(\mathbf{A}_j)\right]\right\} \\ &= \sum_{j=0}^{m} \pi_0^{(j)}(\mathbf{X}^{(j)}, \mathbf{A}^{(j-1)}) I_{\mathbf{x}^{(j)}}(\mathbf{X}^{(j)}) \left\{\mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\mathbf{A}_j, \text{pre}_{\mathbf{T}}(\mathbf{A}_j)\right] - \mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\text{pre}_{\mathbf{T}}(\mathbf{A}_j)\right]\right\} \\ &= \sum_{k=0}^{m} \pi_0^{(k)} I_{\mathbf{x}^{(k)}}(\mathbf{X}^{(k)}) \left\{\mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\mathbf{A}^{(k)}, \mathbf{X}^{(k)}\right] - \mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\mathbf{A}^{(k-1)}, \mathbf{X}^{(k)}\right]\right\}.\end{aligned}$$

To witness the Eq. (A.21), we have to show that the following equations hold for $k = m, \ldots, 0$:

$$\overline{\mu}_0^{k+1} = \mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\mathbf{A}^{(k)}, \mathbf{X}^{(k)}\right] \tag{A.23}$$

$$\mu_0^{k} = \mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\mathbf{A}^{(k-1)}, \mathbf{X}^{(k)}\right]. \tag{A.24}$$

We will prove this by induction. To witness the base case $k = m$:

$$\mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\mathbf{A}^{(m)}, \mathbf{X}^{(m)}\right] = \mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\mathbf{T}\right] = I_{\mathbf{y}}(\mathbf{Y}) = \overline{\mu}_0^{m+1}$$

since $\mathbf{A} \subseteq \mathbf{T}$ by its definition. Also,

$$\mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\mathbf{A}^{(m-1)}, \mathbf{X}^{(m)}\right] = \mathbb{E}\left[I_{\mathbf{y}}(\mathbf{Y})|\mathbf{A}^{(m-1)}, \mathbf{X}^{(m)}\right] = \mu_0^{m},$$

where the first equality holds by Eq. (A.19).

For the induction step, assume that Eqs. (A.23, A.24) hold for $k$. Then

$$\begin{aligned}\mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\mathbf{A}_{k-1}, \mathbf{A}^{(k-2)}, \mathbf{X}^{(k-1)}\right] &= \mathbb{E}_{P_\pi}\left[\mathbb{E}_{I_{\mathbf{y}}(\mathbf{Y})|\mathbf{A}^{(k-1)}, \mathbf{X}^{(k)}}\left[P_\pi\right]|\mathbf{A}_{k-1}, \mathbf{A}^{(k-2)}, \mathbf{X}^{(k-1)}\right] \\ &= \mathbb{E}_{P_\pi}\left[\mu_0^{k}(\mathbf{A}^{(k-1)}, \mathbf{X}^{(k)})|\mathbf{A}_{k-1}, \mathbf{A}^{(k-2)}, \mathbf{X}^{(k-1)}\right] \\ &= \mu_0^{k}(\mathbf{x}_k, \mathbf{A}^{(k-1)}, \mathbf{X}^{(k-1)}) \\ &= \overline{\mu}_0^{k}(\mathbf{x}_k, \mathbf{A}^{(k-1)}, \mathbf{X}^{(k-1)}),\end{aligned} \tag{A.25}$$

where the second equality holds by the induction hypothesis, the third equality comes from the expectation over $P_\pi(\mathbf{x}_k|\mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)}) = I_{\mathbf{x}_k}(\mathbf{X}_k)$, and the last equality holds since $\overline{\mu}_0^{k}$ can be derived by fixing $\mathbf{X}_k$ as $\mathbf{x}_k$ from $\mu_0^{k}$.

Also,

$$\begin{aligned}\mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\mathbf{A}^{(k-2)}, \mathbf{X}^{(k-1)}\right] &= \mathbb{E}_{P_\pi}\left[\mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\mathbf{A}^{(k-1)}, \mathbf{X}^{(k-1)}\right]|\mathbf{A}^{(k-2)}, \mathbf{X}^{(k-1)}\right] \\ &= \mathbb{E}_{P_\pi}\left[\overline{\mu}_0^{k}(\mathbf{x}_k, \mathbf{A}^{(k-1)}, \mathbf{X}^{(k-1)})|\mathbf{A}^{(k-2)}, \mathbf{X}^{(k-1)}\right] \\ &= \mathbb{E}_{P}\left[\overline{\mu}_0^{k}(\mathbf{x}_k, \mathbf{A}^{(k-1)}, \mathbf{X}^{(k-1)})|\mathbf{A}^{(k-2)}, \mathbf{X}^{(k-1)}\right] \\ &= \mu_0^{k-1},\end{aligned}$$

where the second equality holds by Eq. (A.25), the third equality holds by Eq. (A.19), and the last equality by the Def. of $\mu_0^{k-1}$. We conclude that Eqs. (A.23, A.24) hold for $k = m, \ldots, 0$.

Therefore, we can rewrite the influence function as the following:

$$\phi = \sum_{k=0}^{m} \pi_0^{(k)} I_{\mathbf{x}^{(k)}}(\mathbf{X}^{(k)}) \left\{\overline{\mu}_0^{k+1} - \mu_0^{k}\right\}.$$

Now, we will derive the UIF. Note that

$$\phi = \overline{\mu}_0^{1} - \mu_0^{0} + \sum_{k=1}^{m} \pi_0^{(k)} I_{\mathbf{x}^{(k)}}(\mathbf{X}^{(k)}) \left\{\overline{\mu}_0^{k+1} - \mu_0^{k}\right\}.$$

Then by Lemma A.6, $\mu_0^0 = \psi$, which implies that the UIF is given by

$$\mathcal{V} = \overline{\mu}_0^{1} + \sum_{k=1}^{m} \pi_0^{(k)} I_{\mathbf{x}^{(k)}}(\mathbf{X}^{(k)}) \left\{\overline{\mu}_0^{k+1} - \mu_0^{k}\right\}.$$

Using the fact that the IF $\phi$ has a mean-zero property and $\mu_0^0 = \psi$, we can witness the second statement. This completes the proof. □

## Proof for Lemma 4

**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 $\mathbf{A}_j \equiv An(\mathbf{D}_j)_{G(\mathbf{S}_i)}$.*

*Proof.* The proof for Lemma 4 follows from Lemma A.8. □

We first establish notations. For the notational convenience, we will denote $\mathbf{S}$ for any $\mathbf{S}_i$, a $C$-component on $G$, and $\mathbf{D}$ for $\mathbf{D}_j \subseteq \mathbf{S}_i$, a $C$-component on $G(An(\mathbf{V}\setminus\mathbf{X}))$. We will define the *round index* $r$ of MCOMPILE algorithm as the number of recursion in running MCOMPILE. We use $\mathbf{S}_r$ for the $C$-component of $G(\mathbf{A}_{r-1})$ containing $\mathbf{D}$ (where $\mathbf{A}_{-1} = \mathbf{V}$), and $\mathbf{A}_r \equiv An(\mathbf{D})_{G(\mathbf{S}_r)}$. Note $\mathbf{S}_0$ is a $C$-component on $G$ containing $\mathbf{D}$. Let $\mathbf{A}_r = \{A_{r,1}, A_{r,2}\cdots, A_{r,m_r}\}$ where $A_{r,1} \prec A_{r,2} \prec \cdots \prec A_{r,m_r}$ for all $r = 0, 1, 2, \cdots$. At $r$th round, let $Q[\mathbf{A}_r] = \mathcal{A}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=1}^{n_r})$, where $\mathcal{A}^r$ denotes an arithmetic operator; $\mathcal{M}_{r,\ell_r}$ are mSBD operators such that $\mathcal{M}_{r,\ell_r} = \mathcal{M}[\mathbf{y}_{r,\ell_r} \mid \mathbf{x}_{r,\ell_r}; \mathbf{z}_{r,\ell_r}]$.

We first prove a background result:

**Lemma A.7.** *Let $\mathbf{A}_r = \{A_{r,1}, A_{r,2}, \cdots, A_{r,m_r}\}$ where $A_{r,1} \prec \cdots \prec A_{r,m_r}$ for any $r$. Then, (1) $\mathbf{A_r}^{\geq|\mathbf{S}_{r+1}|+1} = \emptyset$ for any $r = 0, 1, \cdots$; (2) $A_{r,m_r} = A_{r,|\mathbf{S}_{r+1}|} \in \mathbf{D}$; and (3) $A_{r,m_r} = A_{0,m_0}$.*

*Proof.* Note, for any $r = 0, 1, 2, \cdots,$; $\mathbf{A}_r = An(\mathbf{D})_{G(\mathbf{S}_r)}$, and $\mathbf{D} \subseteq \mathbf{S}_{r+1}$ by its definition ($\mathbf{S}_r$ is a $C$-component on $G(\mathbf{A}_r)$ containing $\mathbf{D}$). Note $\mathbf{A_r}^{\geq|\mathbf{S}_{r+1}|+1} = \emptyset$; Otherwise, it means there exists a variable $A_{r,|\mathbf{S}_{r+1}|+1} \in \mathbf{A}_r$. Notice $A_{r,|\mathbf{S}_{r+1}|+1} \notin \mathbf{D}$ since $\mathbf{S}_{r+1}$ is a set containing $\mathbf{D}$. Since $\mathbf{A}_r$ is an ancestral set of $\mathbf{D}$, this implies that $A_{r,|\mathbf{S}|_{r+1}+1}$ is also in the ancestral set of $\mathbf{D}$. However, for any $\mathbf{A}_{r,j} \in \mathbf{S}_{r+1}$ (containing $\mathbf{D}$), $\mathbf{A}_{r,j} \prec \mathbf{A}_{r,|\mathbf{S}_{r+1}|+1}$. This is a contradiction to the setting where $\mathbf{A}_r = An(\mathbf{D})_{G(\mathbf{S}_r)}$. Therefore, $\mathbf{A_r}^{\geq|\mathbf{S}|_{r+1}+1} = \emptyset$.

Now, consider $A_{r,|\mathbf{S}_{r+1}|} \in \mathbf{S}_{r+1}$. Suppose $A_{r,|\mathbf{S}_{r+1}|} \notin \mathbf{D}$. Then, there exists $\mathbf{A}_{r,j}$ for $j < |\mathbf{S}_{r+1}|$ that $\mathbf{A}_{r,j} \in \mathbf{D}$ and $\mathbf{A}_{r,j} \prec \mathbf{A}_{r,|\mathbf{S}_{r+1}|}$. This contradicts with the setting that $\mathbf{A}_r = An(\mathbf{D})_{G(\mathbf{S}_r)}$. Therefore, $A_{r,|\mathbf{S}_{r+1}|} \in \mathbf{D}$. That is, for any $r$ and $\mathbf{A}_r = \{A_{r,1}, A_{r,2}, \cdots, A_{r,m_r}\}$, $A_{r,m_r} = A_{r,|\mathbf{S}_{r+1}|} \in \mathbf{D}$. In other words, for any $r$, $\mathbf{A_r}^{\geq|\mathbf{S}_r|+1} = \emptyset$ and $\mathbf{A_r}^{\geq j} \neq \emptyset$ for $j \leq |\mathbf{S}_r|$.

We now see $A_{r,m_r} = A_{0,m_0}$ for any $r$. Notice $A_{0,m_0}$ is a descendent node containing $\mathbf{D}$ in $\mathbf{A}_0$. That is, $A_{0,m_0} \in \mathbf{D}$. This implies $A_{0,m_0} \in \mathbf{A}_r$, since $\mathbf{A}_r$ contains $\mathbf{D}$. Note $A_{0,m_0} \in \mathbf{S}_{r+1}$ since $\mathbf{S}_{r+1}$ contains $\mathbf{D}$. If $A_{r,m_r} \neq A_{0,m_0}$, then $A_{0,m_0} \prec A_{r,m_r}$ by the definition of $A_{r,m_r}$. This contradicts that $A_{0,m_0}$ is a descendent node in the superset $\mathbf{A}_0$. Therefore, $A_{r,m_r} = A_{0,m_0}$. □

**Lemma A.8 (Primary mSBD operator).** *$\mathcal{A}^j(\{\mathcal{M}_\ell^j\}_{\ell=1}^{m_j}) = \sum_{\cdot} \mathcal{M}_1^j \mathcal{B}^j(\{\mathcal{M}_\ell^j\}_{\ell=2}^{m_j})$, where $\mathcal{M}_1^j$ is a primary mSBD operator $\mathcal{M}_1^j = \mathcal{M}[\mathbf{a}_j \mid Pa(\mathbf{s}_i)\setminus\mathbf{s}_i; \mathbf{s}_i\setminus\mathbf{a}_j]$; $\mathcal{M}_\ell^j = \mathcal{M}[\mathbf{y}_{j,\ell} \mid \mathbf{x}_{j,\ell}; \mathbf{z}_{j,\ell}]$ for $\ell \geq 2$ are mSBD operators such that $A_{m_0} \notin \mathbf{Y}_{j,\ell}$; $\mathcal{M}_\ell^j$ for $\ell \geq 2$ is obtained by marginalization of $\mathcal{M}_1^j$ by Lemma 2; and $\mathcal{B}$ an arithmetic combination operator that does not contain $\mathcal{M}_1^j$ as its argument.*

*Proof.* We first make an inductive hypothesis at $r$th round: At $r$th round, suppose $Q[\mathbf{A}_r] = \mathcal{A}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=1}^{n_r}) = \sum_{\cdot} \mathcal{M}_1 \mathcal{B}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=2}^{n_r})$, where $\mathcal{M}_1 = \mathcal{M}_{r,1}$; $\mathcal{M}_{r,\ell_r}$ for $\ell_r \geq 2$ are mSBD operators such that $A_{0,m_0} \notin \mathbf{Y}_{r,\ell_r}$; $\mathcal{M}_{r,\ell_r}$ for $\ell_r \geq 2$ is obtained by marginalization of $\mathcal{M}_1^j$ by Lemma 2; and $\mathcal{B}^r$ an arithmetic operator, which does not contain $\mathcal{M}_1$ as its argument.

Then, at $r+1$'th round,

$$Q[\mathbf{A}_{r+1}] \tag{A.26}$$

$$= \sum_{\mathbf{S}_{r+1}\setminus\mathbf{A}_{r+1}} \prod_{A_{r,j}\in\mathbf{S}_{r+1}} \frac{\sum_{\mathbf{A_r}^{\geq j+1}} Q[\mathbf{A}_r]}{\sum_{\mathbf{A_r}^{\geq j}} Q[\mathbf{A}_r]}, \text{ by MCOMPILE algorithm}$$

$$= \sum_{\mathbf{S}_{r+1}\setminus\mathbf{A}_{r+1}} Q[\mathbf{A}_r] \frac{1}{\sum_{A_{r,|\mathbf{S}_{r+1}|}} Q[\mathbf{A}_r]} \prod_{A_{r,j}\in\mathbf{S}_{r+1}\setminus\{A_{r,|\mathbf{S}_{r+1}|}\}} \frac{\sum_{\mathbf{A_r}^{\geq j+1}} Q[\mathbf{A}_r]}{\sum_{\mathbf{A_r}^{\geq j}} Q[\mathbf{A}_r]} \tag{A.27}$$

$$= \sum_{\mathbf{S}_{r+1}\setminus\mathbf{A}_{r+1}} \sum_{\cdot} \mathcal{M}_1 \mathcal{B}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=2}^{n_r}) \frac{1}{\sum_{A_{r,|\mathbf{S}_{r+1}|}} \sum_{\cdot} \mathcal{M}_1 \mathcal{B}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=2}^{n_r})} \prod_{A_{r,j}\in\mathbf{S}_{r+1}\setminus\{A_{r,|\mathbf{S}_{r+1}|}\}} \frac{\sum_{\mathbf{A_r}^{\geq j+1}} \sum_{\cdot} \mathcal{M}_1 \mathcal{B}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=2}^{n_r})}{\sum_{\mathbf{A_r}^{\geq j}} \sum_{\cdot} \mathcal{M}_1 \mathcal{B}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=2}^{n_r})}, \tag{A.28}$$

where Eq. (A.27) holds since $A_{r,|\mathbf{S}_{r+1}|+1} = \emptyset$, by Lemma A.7; Eq. (A.28) is by the inductive hypothesis at $r$'th round. For any $j = 1, 2, \cdots, |\mathbf{S}_{r+1}|$,

$$\begin{aligned}
&\sum_{\mathbf{A_r}^{\geq j}} \sum_{\cdot} \mathcal{M}_1 \mathcal{B}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=2}^{n_r}) \\
&= \sum_{\cdot} \sum_{\mathbf{A_r}^{\geq j}} \mathcal{M}_1 \mathcal{B}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=2}^{n_r}), \\
&= \sum_{\cdot} \sum_{\mathbf{A_r}^{\geq j} \setminus \{A_{r,m_r}\}} \sum_{A_{r,m_r}} \mathcal{M}_1 \mathcal{B}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=2}^{n_r}) \\
&= \sum_{\cdot} \sum_{\mathbf{A_r}^{\geq j} \setminus \{A_{r,m_r}\}} \sum_{A_{0,m_0}} \mathcal{M}_1 \mathcal{B}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=2}^{n_r}), \text{ by Lemma A.7} \\
&= \sum_{\cdot} \left( \sum_{A_{0,m_0}} \mathcal{M}_1 \right) \sum_{\mathbf{A_r}^{\geq j} \setminus \{A_{0,m_0}\}} \mathcal{B}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=2}^{n_r}), \text{ by the inductive hypothesis at } r\text{th round} \\
&= \sum_{\cdot} \mathcal{M}_{r,0} \sum_{\mathbf{A_r}^{\geq j} \setminus \{A_{0,m_0}\}} \mathcal{B}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=2}^{n_r}), \text{ for } \mathcal{M}_{r,0} \text{ a mSBD operator such that } A_{0,m_0} \notin \mathbf{Y}_{r,0} \\
&\equiv \mathcal{C}_j(-\mathcal{M}_1), \text{ where } \mathcal{C}_{\cdot}(\cdot) \text{ a mSBD operator that does not have } \mathcal{M}_1 \text{ as its arguments.}
\end{aligned}$$

Then,

$$\begin{aligned}
Q[\mathbf{A}_{r+1}] &= \sum_{\mathbf{S}_{r+1} \setminus \mathbf{A}_{r+1}} \sum_{\cdot} \mathcal{M}_1 \mathcal{B}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=2}^{n_r}) \frac{1}{\mathcal{C}_{|\mathbf{S}_{r+1}|}(-\mathcal{M}_1)} \prod_{A_{r,j} \in \mathbf{S}_{r+1} \setminus \{A_{r,|\mathbf{S}_{r+1}|}\}} \frac{\mathcal{C}_{j+1}(-\mathcal{M}_1)}{\mathcal{C}_j(-\mathcal{M}_1)} \quad \text{(A.29)} \\
&\equiv \mathcal{A}^{r+1}(\{\mathcal{M}_{r+1,\ell_{r+1}}\}_{\ell_{r+1}=1}^{n_{r+1}}),
\end{aligned}$$

for some mSBD operators $\mathcal{M}_{r+1,\ell_{r+1}}$ composing Eq. (A.29), and $\mathcal{A}^{r+1}(\cdot)$ an arithmetic combination operator mapping $\{\mathcal{M}_{r+1,\ell_{r+1}}\}_{\ell_{r+1}=1}^{n_{r+1}}$ to Eq. (A.29). Without loss of generality, we can set $\mathcal{M}_{r+1,1} = \mathcal{M}_1$, since Eq. (A.29) contains $\mathcal{M}_1$.

Let

$$\mathcal{B}^{r+1}(\{\mathcal{M}_{r+1,\ell_{r+1}}\}_{\ell_{r+1}=2}^{n_{r+1}}) \equiv \mathcal{B}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=2}^{n_r}) \frac{1}{\mathcal{C}_{|\mathbf{S}_{r+1}|}(-\mathcal{M}_1)} \prod_{A_{r,j} \in \mathbf{S}_{r+1} \setminus \{A_{r,|\mathbf{S}_{r+1}|}\}} \frac{\mathcal{C}_{j+1}(-\mathcal{M}_1)}{\mathcal{C}_j(-\mathcal{M}_1)},$$

where $\mathcal{M}_{r+1,\ell_{r+1}}$ are mSBD operators composing $\mathcal{B}^{r+1}$. Note $\mathcal{M}_{r+1,\ell_{r+1}}$ for $\ell_{r+1} \geq 2$ are mSBD operators such that $A_{0,m_0} \notin \mathbf{Y}_{r+1,\ell_{r+1}}$ by the inductive hypothesis made for $Q[\mathbf{A}_r]$. Then, we can witness that that $Q[\mathbf{A}_{r+1}] = \mathcal{A}^{r+1}(\{\mathcal{M}_{r+1,\ell_r}\}_{\ell_{r+1}=1}^{n_{r+1}}) = \sum_{\cdot} \mathcal{M}_1 \mathcal{B}^{r+1}(\{\mathcal{M}_{r+1,\ell_{r+1}}\}_{\ell_{r+1}=2}^{n_{r+1}})$, where $\mathcal{M}_1 = \mathcal{M}_{r+1,1}$; $\mathcal{M}_{r+1,\ell_{r+1}}$ for $\ell_{r+1} \geq 2$ are mSBD operators such that $A_{0,m_0} \notin \mathbf{Y}_{r+1,\ell_{r+1}}$. Specifically, $\mathcal{M}_{r+1,\ell_{r+1}}$ for $\ell_{r+1} \geq 2$ is either the same as $\mathcal{M}_{r,\ell_r}$ for some $\ell_r$, or given by $\mathcal{M}_{r+1,\ell_{r+1}} = \sum_{\mathbf{A_r}^{\geq j}} \mathcal{M}_{r,\ell_r}$ for some $\ell_r$ and $j$, if $\mathbf{A_r}^{\geq j} = De(\mathbf{A_r}^{\geq j})_{G[\mathbf{Y}_{r,\ell_r}]}$ or $\mathbf{A_r}^{\geq j} = An(\mathbf{A_r}^{\geq j})_{G[\mathbf{Y}_{r,\ell_r}]}$ (Lemma 2). In either cases, $\mathcal{M}_{r+1,\ell_{r+1}}$ for $\ell_{r+1}$ is obtained by marginalization of $\mathcal{M}_1^j$ by Lemma 2, by the inductive hypothesis. We note $\mathcal{M}_{r+1,\ell_{r+1}}$ is a mSBD operator distinct to $\mathcal{M}_1$, since the marginalization $\mathbf{A_r}^{\geq j}$ includes $A_{0,m_0}$, by Lemma A.7. Finally, $\mathcal{B}^{r+1}$ an arithmetic operator, which does not contain $\mathcal{M}_1$ as its argument. Therefore, the inductive hypothesis at $r+1$'th round is also satisfied.

We now check the initial condition at $r = 0$ and $r = 1$. For $r = 0$, $Q[\mathbf{A}_r] = Q[\mathbf{A}_0] = \mathcal{M}_1 = \mathcal{M}[\mathbf{a}_0 \mid Pa(\mathbf{s}_0) \setminus \mathbf{s}_0; \mathbf{s}_0 \setminus \mathbf{a}_0]$ by Lemma 1. For $r = 1$,

$$\begin{aligned}
Q[\mathbf{A}_1] &= \sum_{\mathbf{S}_1 \setminus \mathbf{A}_1} Q[\mathbf{A}_0] \frac{1}{\sum_{A_{0,|\mathbf{S}_0|}} Q[\mathbf{A}_0]} \prod_{j=1}^{|\mathbf{S}_0|-1} \frac{\sum_{\mathbf{A_0}^{\geq j+1}} Q[\mathbf{A}_0]}{\sum_{\mathbf{A_0}^{\geq j}} Q[\mathbf{A}_0]} \\
&= \sum_{\mathbf{S}_1 \setminus \mathbf{A}_1} \mathcal{M}_1 \frac{1}{\sum_{A_{0,|\mathbf{S}_0|}} Q[\mathbf{A}_0]} \prod_{j=1}^{|\mathbf{S}_0|-1} \frac{\sum_{\mathbf{A_0}^{\geq j+1}} Q[\mathbf{A}_0]}{\sum_{\mathbf{A_0}^{\geq j}} Q[\mathbf{A}_0]}.
\end{aligned}$$

We note $\sum_{\mathbf{A_0}^{\geq j}} Q[\mathbf{A}_0] = \sum_{\mathbf{A_0}^{\geq j}} \mathcal{M}_1$ for $j = 1, 2, \cdots, |\mathbf{S}_0|$ not only marginalizes out $A_{0,m_0} = A_{0,|\mathbf{S}_0|}$ (by Lemma A.7), but also renders a mSBD operators distinct to $\mathcal{M}_1$, by Lemma 2, since $\mathbf{A_0}^{\geq j} = De(\mathbf{A_0}^{\geq j})_{G[\mathbf{A}_0]}$. Therefore, $\sum_{\mathbf{A_0}^{\geq j}} \mathcal{M}_1$ yields mSBD operators $\mathcal{M}_\ell$ (for $\ell \geq 2$) such that $A_{0,m_0} \notin \mathbf{Y}_\ell$. This implies that, for

$$\mathcal{B}^1(\{\mathcal{M}_{\ell_1}\}) \equiv \frac{1}{\sum_{A_{0,|\mathbf{S}_0|}} Q[\mathbf{A}_0]} \prod_{j=1}^{|\mathbf{S}_0|-1} \frac{\sum_{\mathbf{A_0} \geq j+1} Q[\mathbf{A}_0]}{\sum_{\mathbf{A_0} \geq j} Q[\mathbf{A}_0]},$$

$\mathcal{B}^1$ does not have mSBD operators $\mathcal{M}_\ell$ such that $A_{0,m_0} \in \mathbf{Y}_\ell$ (because it is marginalized out). Therefore, the inductive hypothesis is true for $r=0, r=1$. Combining for the general $r$'th round, we conclude that the inductive hypothesis is true.

Therefore, for any $r$, $Q[\mathbf{A}_r] = \mathcal{A}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=1}^{n_r}) = \sum_{\cdot} \mathcal{M}_1 \mathcal{B}^r(\{\mathcal{M}_{r,\ell_r}\}_{\ell_r=2}^{n_r})$, where $\mathcal{M}_1 = \mathcal{M}_{r,1}$; $\mathcal{M}_{r,\ell_r}$ for $\ell_r \geq 2$ are obtained by marginalization of $\mathcal{M}_1$ such that $A_{0,m_0} \notin \mathbf{Y}_{r,\ell_r}$; and $\mathcal{B}^r$ an arithmetic operator, which does not contain $\mathcal{M}_1$ as its argument. This completes the proof, since $Q[\mathbf{D}] = Q[\mathbf{A}_{r'}]$ for some $r'$ (by the return condition of MCOMPILE).

□

## Proof for Lemma 5

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

*Proof.* Consider $Q[\mathbf{D}_j] = \mathcal{A}^j(\{\mathcal{M}^j_\ell\})$. In this proof, we denote $\mathcal{M}(P_t)$ for representing a mSBD operator defined on $P_t \equiv P(1+tg)$, a parametric submodel where $t \in \mathbb{R}$ and $g$ a mean-zero bounded random function. Then, the target functional defined on the submodel is given by $\Psi(P_t) = \mathcal{A}^j(\{\mathcal{M}^j_\ell(P_t)\}) = (\mathcal{A}^j \circ \{\mathcal{M}^j_\ell\})(P_t)$ ($\circ$ is a general composition operator between two functional), where $\psi = \Psi(P_0)$. For any functional $f(P)$, let $\nabla_g f \equiv \lim_{t \to 0} \frac{f(P+tPg)-f(P)}{t}|_{t=0} = \frac{\partial}{\partial t} f(P+tPg)|_{t=0}$. Then, by definition, an IF of $Q[\mathbf{D}_j]$ is given by a function $\phi_{Q[\mathbf{D}_j]}$ satisfying $\nabla_g \Psi = \mathbb{E}_P[\phi_{Q[\mathbf{D}_j]} \cdot S_t(\mathbf{V}; t=0)]$, where $\phi$ has mean-zero and finite variance. We have,

$$\begin{aligned}\nabla_g \Psi &= \nabla_g(\mathcal{A}^j \circ \{\mathcal{M}^j_\ell\}) \\ &= \sum_{\ell=1}^{m_j} \nabla_{\nabla_g \mathcal{M}^j_\ell} \mathcal{A}^j \text{ by multivariate chain rule of Gateaux derivative,}\end{aligned}$$

where

$$\gamma_\ell \equiv \nabla_g \mathcal{M}^j_\ell = \mathbb{E}_P\left[\phi_{\mathcal{M}^j_\ell} \cdot S_t(\mathbf{V}; t=0)\right],$$

where $\phi_{\mathcal{M}^j_\ell}$ is an IF of a mSBD operator $\mathcal{M}^j_\ell$, by definition of an IF of mSBD operator.

Then, we can rewrite as $\nabla_g \Psi = \sum_{\ell=1}^{m_j} \nabla_{\gamma_\ell} \mathcal{A}^j$. We note $\nabla_{r_\ell} \mathcal{A}^j \equiv \lim_{t \to 0} \frac{\mathcal{A}^j(\mathcal{M}^j_\ell + t\gamma_\ell) - \mathcal{A}^j(\mathcal{M}^j_\ell)}{t} = \frac{\partial}{\partial t} \mathcal{A}^j(\mathcal{M}^j_\ell + t\gamma_\ell)|_{t=0}$, which could be found by conducting a directional derivative.

If $\mathcal{A}^j$ is not a function of $\mathcal{M}^j_\ell$ (line a.2 of Algo. 2), then $\nabla_{\gamma_\ell} \mathcal{A}^j = 0$, since $\mathcal{A}^j(\mathcal{M}^j_\ell + t\gamma_\ell) = \mathcal{A}^j(\mathcal{M}^j_\ell)$.

If $\mathcal{A}^j = \mathcal{M}^j_\ell$ (line a.3 of Algo. 2), then $\nabla_{\gamma_\ell} \mathcal{A}^j = \gamma_\ell$, since $\mathcal{A}^j(\mathcal{M}^j_\ell + t\gamma_\ell) - \mathcal{A}^j(\mathcal{M}^j_\ell) = t\gamma_\ell$.

If $\mathcal{A}^j = C\mathcal{A}'^j$ (line a.4 of Algo. 2), then $\nabla_{\gamma_\ell} \mathcal{A}^j = C\nabla_{\gamma_\ell} \mathcal{A}'^j$, since $\mathcal{A}^j(\mathcal{M}^j_\ell + t\gamma_\ell) - \mathcal{A}^j(\mathcal{M}^j_\ell) = C\left(\mathcal{A}'^j(\mathcal{M}^j_\ell + t\gamma_\ell) - \mathcal{A}'^j(\mathcal{M}^j_\ell)\right)$.

If $\mathcal{A}^j = \mathcal{A}'^j \mathcal{A}''^j$ (line a.5 of Algo. 2), then $\nabla_{\gamma_\ell} \mathcal{A}^j = \nabla_{\gamma_\ell}\left(\mathcal{A}'^j \mathcal{A}''^j\right) = \mathcal{A}''^j \nabla_{\gamma_\ell} \mathcal{A}'^j + \mathcal{A}'^j \nabla_{\gamma_\ell} \mathcal{A}''^j$. This rule subsumes line a.6 of Algo. 2, when $\mathcal{A}' \leftarrow 1$ and $\mathcal{A}'' \leftarrow 1/\mathcal{A}'$.

If $\mathcal{A}^j$ has a form $\mathcal{A}^j = \sum \mathcal{A}'^j$ (line a.7 of Algo. 2),

$$\begin{aligned}\nabla_{r_\ell} \mathcal{A}^j &\equiv \lim_{t \to 0} \frac{\mathcal{A}^j(\mathcal{M}^j_\ell + t\gamma_\ell) - \mathcal{A}^j(\mathcal{M}^j_\ell)}{t} \\ &= \lim_{t \to 0} \sum \frac{\mathcal{A}'^j(\mathcal{M}^j_\ell + t\gamma_\ell) - \mathcal{A}'^j(\mathcal{M}^j_\ell)}{t} \\ &= \sum \lim_{t \to 0} \frac{\mathcal{A}'^j(\mathcal{M}^j_\ell + t\gamma_\ell) - \mathcal{A}'^j(\mathcal{M}^j_\ell)}{t} \text{ by Dominated Convergence Theorem.} \quad \text{(A.30)} \\ &= \sum \nabla_{\gamma_\ell} \mathcal{A}'^j. \quad \text{(A.31)}\end{aligned}$$

Since an arithmetic combination operator $\mathcal{A}$ composes of multiplication/marginalization/division of $\mathcal{M}$, applying product/interchange/quotient rules discussed above are sufficient.

Note that $\sum \gamma_\ell \cdot f(\{\mathcal{M}\})$ for some function $f$ and general marginalization $\sum$ could be written as $\mathbb{E}_P\left[\left(\sum \phi_{\mathcal{M}^j_\ell} \cdot f(\{\mathcal{M}\})\right) \cdot S_t(\mathbf{V}; t=0)\right]$, using the definition of $\gamma_\ell$ (we call this procedure as '*Extraction*' for this proof). Then, one can see that $\textsc{FindH}(\mathcal{A}^j(\{\mathcal{M}_\ell^j\}, \mathcal{M}_\ell^j)$ computes $\nabla_{\gamma_\ell}$ and conducts the *extraction* procedure. This implies that an IF of $Q[\mathbf{D}_j]$ can be obtained based on

$$\begin{aligned}
\nabla_g \Psi &= \nabla_g(\mathcal{A}^j \circ \{\mathcal{M}_\ell^j\}) \\
&= \sum_{\ell=1}^{m_j} \nabla_{\nabla_g \mathcal{M}_\ell^j} \mathcal{A}^j \\
&= \mathbb{E}_P \left[ \left( \sum_{\ell=1}^{m_j} \textsc{ComponentUIF}(\mathcal{A}^j, \mathcal{M}_\ell^j) \right) \cdot S_t(\mathbf{V}; t=0) \right].
\end{aligned}$$

Notice $\mathbb{E}_P\left[\sum_{\ell=1}^{m_j} \textsc{ComponentUIF}(\mathcal{A}^j, \mathcal{M}_\ell^j)\right] = 0$, since $\sum_{\ell=1}^{m_j} \textsc{ComponentUIF}(\mathcal{A}^j, \mathcal{M}_\ell^j)$ is a linear combination of IFs of mSBD operators, which has mean zero, and finite variance under general positivity assumption. This completes the proof. □

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

*Proof.* Note

$$\begin{aligned}
\nabla_g \Psi &= \nabla_g(\mathcal{A}^j \circ \{\mathcal{M}_\ell^j\}) \\
&= \sum_{\ell=1}^{m_j} \nabla_{\nabla_g \mathcal{M}_\ell^j} \mathcal{A}^j \text{ by multivariate chain rule of Gateaux derivative,}
\end{aligned}$$

where

$$\gamma_\ell \equiv \nabla_g \mathcal{M}_\ell^j = \mathbb{E}_P\left[\phi_{\mathcal{M}_\ell^j} \cdot S_t(\mathbf{V}; t=0)\right],$$

where $\phi_{\mathcal{M}_\ell^j}$ an IF of a mSBD operator $\mathcal{M}_\ell^j$, by definition of an IF of mSBD operator. Note

$$\nabla_{\gamma_\ell} \mathcal{A}^j \equiv \lim_{t \to 0} \frac{\mathcal{A}^j\left(\mathcal{M}_\ell^j + t\gamma_\ell\right) - \mathcal{A}^j\left(\mathcal{M}_\ell^j\right)}{t}.$$

Since $\mathcal{A}^j$ is an arithmetic combination, with a general positivity assumption, the derivative of $\mathcal{A}^j$, denoted $\nabla \mathcal{A}^j \equiv \frac{\partial}{\partial \mathcal{M}_\ell^j} \mathcal{A}^j$, exists. Since $\mathcal{A}^j$ does not contain marginalization, the directional derivative in the direction $\gamma_\ell$ equals to $\nabla \mathcal{A}^j \cdot \gamma_\ell$ (Marsden, Hoffman et al. 1993, Thm. 6.4.1) (i.e., $\nabla_{\gamma_\ell} \mathcal{A}^j = \gamma_\ell \cdot \nabla \mathcal{A}^j$), we note

$$\nabla_g \Psi = \sum_{\ell=1}^{m_j} \nabla_{\nabla_g \mathcal{M}_\ell^j} \mathcal{A}^j = \sum_{\ell=1}^{m_j} \nabla_{\gamma_\ell} \mathcal{A}^j = \sum_{\ell=1}^{m_j} \gamma_\ell \cdot \nabla \mathcal{A}^j = \sum_{\ell=1}^{m_j} \gamma_\ell \cdot \frac{\partial}{\partial \mathcal{M}_\ell^j} \mathcal{A}^j.$$

By the '*extraction*' procedure, defined in proof of Lemma 5; and the equality $\frac{\partial}{\partial \mathcal{M}_\ell^j} \mathcal{A}^j(\{\mathcal{M}_\ell^j\}) = \frac{\partial}{\partial \mu_{\mathcal{M}_\ell^j}} \mathcal{A}^j(\{\mu_{\mathcal{M}_\ell^j}\})$, we note

$$\begin{aligned}
\nabla_g \Psi &= \sum_{\ell=1}^{m_j} \mathbb{E}_P\left[\phi_{\mathcal{M}_\ell^j} \frac{\partial}{\partial \mu_{\mathcal{M}_\ell^j}} \mathcal{A}^j(\{\mu_{\mathcal{M}_\ell^j}\}) \cdot S_t(\mathbf{V}; t=0)\right] \\
&= \mathbb{E}_P\left[\left(\sum_{\ell=1}^{m_j} \phi_{\mathcal{M}_\ell^j} \frac{\partial}{\partial \mu_{\mathcal{M}_\ell^j}} \mathcal{A}^j(\{\mu_{\mathcal{M}_\ell^j}\})\right) \cdot S_t(\mathbf{V}; t=0)\right],
\end{aligned}$$

implying, from the proof of Lemma 5, that

$$\textsc{CompoUIF}(\mathcal{A}^j, \mathcal{M}_\ell^j) = \phi_{\mathcal{M}_\ell^j} \frac{\partial}{\partial \mu_{\mathcal{M}_\ell^j}} \mathcal{A}^j(\{\mu_{\mathcal{M}_\ell^j}\}) = (\mathcal{V}_{\mathcal{M}_\ell^j} - \mu_{\mathcal{M}_\ell^j}) \frac{\partial}{\partial \mu_{\mathcal{M}_\ell^j}} \mathcal{A}^j(\{\mu_{\mathcal{M}_\ell^j}\}).$$

□

## Proof for Theorem 2

**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 $\psi$ 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*

$$\begin{aligned}\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})} &= \sum_{\mathbf{d}\setminus\mathbf{y}} \mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1^1}, \{\mu_{\mathcal{M}_l^1}\}_{\ell=2}^{m_1}) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_p}) \\ &+ \sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{\ell=2}^{m_1} h_{\mathcal{A}^1,\mathcal{M}_\ell^1} \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_p}) \\ &+ \sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j=2}^{k_d} \left( \sum_{\ell=1}^{m_j} h_{\mathcal{A}^j,\mathcal{M}_\ell^j} \right) \prod_{\substack{p=1 \\ p\neq j}}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_p}),\end{aligned} \tag{A.32}$$

*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 $h_{\mathcal{A}^j,\mathcal{M}_\ell^j} = \text{ComponentUIF}(\mathcal{A}^j, \mathcal{M}_\ell^j)$.*

*Proof.* Note that the causal effect $P_{\mathbf{x}}(\mathbf{y})$ is given by

$$\Psi(P) \equiv P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{d}\setminus\mathbf{y}} \prod_{j=1}^{k_d} Q[\mathbf{D}_j], \tag{A.33}$$

by line 7 of DML-ID at Algo. 1, where $Q[\mathbf{D}_j] = \text{mCompile}(\mathbf{D}_j, \mathbf{S}_i, Q[\mathbf{S}_i])$ where $\mathbf{S}_i, \mathbf{D}_j, Q[\mathbf{S}_j]$ are defined in line 2,3,5 of Mosaic. Let $\mathcal{A}^j$ denote the arithmetic combination mapping such that $\mathcal{A}^j(\{\mathcal{M}_\ell^j\}_{\ell=1}^{m_j}) = Q[\mathbf{D}_j]$. Note $\mathcal{A}_\mu^j(\{\mu_{\mathcal{M}_\ell^j}\}_{\ell=1}^{m_j}) = \mathcal{A}^j(\{\mathcal{M}_\ell^j\}_{\ell=1}^{m_j})$ by the given setting.

Recall that $P_t \equiv P(1+tg)$ be the parametric-submodel, as defined in Sec. 2. Let $\mathcal{M}(P_t)$ be the mSBD operator defined over the submodel $P_t$. Then, $\frac{\partial}{\partial t}\Psi(P_t)|_{t=0} = \mathbb{E}[\phi_{P_{\mathbf{x}}(\mathbf{y})} S_t(\mathbf{V}; t=0)]$, by the definition of the IF. From the result in Lemma 5,

$$\begin{aligned}\frac{\partial}{\partial t}\Psi(P_t)|_{t=0} &= \nabla_g \sum_{\mathbf{d}\setminus\mathbf{y}} \prod_{j=1}^{k_d} Q[\mathbf{D}_j](P) \\ &= \nabla_g \sum_{\mathbf{d}\setminus\mathbf{y}} \prod_{j=1}^{k_d} (\mathcal{A}^j \circ \{\mathcal{M}_\ell^j(P_t)\}_{\ell=1}^{m_j})(P) \\ &= \sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j=1}^{k_d} \nabla_g \left(\mathcal{A}_k^j \circ \mathcal{M}_k^j\right)(P) \prod_{p\neq j} \mathcal{A}^j(\{\mathcal{M}_\ell^p\}_{\ell=1}^{m_j}) \\ &= \sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j=1}^{k_d} \nabla_{\nabla_g \mathcal{M}_k^j} \mathcal{A}_k^j(\mathcal{M}_k^j) \prod_{p\neq j} \mathcal{A}^j(\{\mathcal{M}_\ell^p\}_{\ell=1}^{m_j}) \\ &= \mathbb{E}_P\left[\left(\sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j=1}^{k_d} \left(\sum_{\ell=1}^{m_j} h_{\mathcal{A}^j,\mathcal{M}_\ell^j}\right) \prod_{p\neq j} \mathcal{A}^j(\{\mathcal{M}_\ell^p\}_{\ell=1}^{m_j})\right) \cdot S_t(\mathbf{V}; t=0)\right],\end{aligned}$$

implying that

$$\phi_{P_{\mathbf{x}}(\mathbf{y})} = \sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j=1}^{k_d} \left(\sum_{\ell=1}^{m_j} h_{\mathcal{A}^j,\mathcal{M}_\ell^j}\right) \prod_{p\neq j} \mathcal{A}^j(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_j}).$$

Notice that $\mathbb{E}_P\left[\phi_{P_{\mathbf{x}}(\mathbf{y})}\right] = 0$, since $\phi_{P_{\mathbf{x}}(\mathbf{y})}$ is expressed as a linear combination of IFs of mSBD operators, which has zero mean. Under a general positivity assumption, a finite variance of $\phi_{P_{\mathbf{x}}(\mathbf{y})}$ is guaranteed by finite variances of IFs of mSBD operators.

We now consider the primary mSBD operators. Note that Lemma. A.8 implies that any arithmetic operator $\mathcal{A}^j(\{\mathcal{M}_\ell\}_{\ell=1}^{m_j})$ could be written as $\mathcal{A}^j(\{\mathcal{M}_\ell\}_{\ell=1}^{m_j}) = \sum \mathcal{M}_1^j \mathcal{B}^j(\{\mathcal{M}_\ell\}_{\ell=2}^{m_j})$ for some arithmetic combination operator $\mathcal{B}$ such that its argument does not contain $\mathcal{M}_1^j$.

We simplify the notation as $\mathcal{A} = \mathcal{A}^j$; $\mathcal{B} = \mathcal{B}^j$; and $\mathcal{M}_1 = \mathcal{M}_1^j$. Let $\mathcal{A}'(\mathcal{M}_1, \{\mathcal{M}_\ell\}_{\ell=2}^m) \equiv \mathcal{M}_1\mathcal{B}(\{\mathcal{M}_\ell\}_{\ell=2}^m)$. Note $\mathcal{A}(\{\mathcal{M}_\ell\}_{\ell=1}^m) = \mathcal{A}(\mathcal{M}_1, \{\mathcal{M}_\ell\}_{\ell=2}^m) = \sum \mathcal{A}'(\mathcal{M}_1, \{\mathcal{M}_\ell\}_{\ell=2}^m)$. Then, by running ComputeUIF, one can see that

$$\begin{aligned}
h_{\mathcal{A},\mathcal{M}_1} &= \text{FindH}(\mathcal{A}, \mathcal{M}_1) \\
&= \sum \text{FindH}(\mathcal{M}_1\mathcal{B}, \mathcal{M}_1), \text{ by Lemma A.8} \\
&= \sum \mathcal{B}(\{\mathcal{M}_\ell\}_{\ell=2}^m)\text{FindH}(\mathcal{M}_1, \mathcal{M}_1) \\
&= \sum \mathcal{B}(\{\mathcal{M}_\ell\}_{\ell=2}^m)\phi_{\mathcal{M}_1} \\
&= \sum \mathcal{A}'(\phi_{\mathcal{M}_1}, \{\mathcal{M}_\ell\}_{\ell=2}^m) \\
&= \mathcal{A}(\phi_{\mathcal{M}_1}, \{\mathcal{M}_\ell\}_{\ell=2}^m).
\end{aligned}$$

Using $\phi_{\mathcal{M}_1} = \mathcal{V}_{\mathcal{M}_1} - \mu_{\mathcal{M}}$, one can rewrite it as $h_{\mathcal{A},\mathcal{M}_1} = \mathcal{A}(\mathcal{V}_{\mathcal{M}_1}, \{\mathcal{M}_\ell\}_{\ell=2}^m) - \mathcal{A}(\mu_{\mathcal{M}_1}, \{\mathcal{M}_\ell\}_{\ell=2}^m)$. By $\mathcal{M} = \mu_{\mathcal{M}}$, we have

$$h_{\mathcal{A},\mathcal{M}_1} = \mathcal{A}(\mathcal{V}_{\mathcal{M}_1}, \{\mu_{\mathcal{M}_\ell}\}_{\ell=2}^m) - \mathcal{A}(\{\mu_{\mathcal{M}_\ell}\}_{\ell=2}^m).$$

We now derive the UIF. Consider a following representation for an IF.

$$\begin{aligned}
&\phi_{P_{\mathbf{x}}(\mathbf{y})} \\
&= \sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j\neq 1} \left( \sum_{\ell=1}^{m_j} h_{\mathcal{A}^j,\mathcal{M}_\ell^j} \right) \prod_{p\neq j} \mathcal{A}^p(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_p}) + \sum_{\mathbf{d}\setminus\mathbf{y}} \left( \sum_{\ell=1}^{m_1} h_{\mathcal{A}^1,\mathcal{M}_\ell^1} \right) \prod_{p\neq 1} \mathcal{A}^p(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_p}) \\
&= \sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j\neq 1} \left( \sum_{\ell=1}^{m_j} h_{\mathcal{A}^j,\mathcal{M}_\ell^j} \right) \prod_{p\neq j} \mathcal{A}^p(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_p}) + \sum_{\mathbf{d}\setminus\mathbf{y}} \left( h_{\mathcal{A}^1,\mathcal{M}_1^1} + \sum_{\ell=2}^{m_1} h_{\mathcal{A}^1,\mathcal{M}_\ell^1} \right) \prod_{p\neq 1} \mathcal{A}^1(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_1}) \\
&= \sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j\neq 1} \left( \sum_{\ell=1}^{m_j} h_{\mathcal{A}^j,\mathcal{M}_\ell^j} \right) \prod_{p\neq j} \mathcal{A}^j(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_j}) + \sum_{\mathbf{d}\setminus\mathbf{y}} \left( \mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1^1}, \{\mu_{\mathcal{M}_\ell^1}\}_{\ell=2}^m) - \mathcal{A}^1(\{\mu_{\mathcal{M}_\ell^1}\}_{\ell=2}^m) + \sum_{\ell=2}^{m_1} h_{\mathcal{A}^1,\mathcal{M}_\ell^1} \right) \prod_{p\neq 1} \mathcal{A}^p(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_1}) \\
&= \sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j\neq 1} \left( \sum_{\ell=1}^{m_j} h_{\mathcal{A}^j,\mathcal{M}_\ell^j} \right) \prod_{p\neq j} \mathcal{A}^j(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_j}) + \sum_{\mathbf{d}\setminus\mathbf{y}} \left( \mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1^1}, \{\mu_{\mathcal{M}_\ell^1}\}_{\ell=2}^m) + \sum_{\ell=2}^{m_1} h_{\mathcal{A}^r,\mathcal{M}_\ell^r} \right) \prod_{p\neq 1} \mathcal{A}^r(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_1}) - \psi.
\end{aligned}$$

This implies that an UIF is given as

$$\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})} = \sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j\neq 1} \left( \sum_{\ell=1}^{m_j} h_{\mathcal{A}^j,\mathcal{M}_\ell^j} \right) \prod_{p\neq j} \mathcal{A}^j(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_j}) + \sum_{\mathbf{d}\setminus\mathbf{y}} \left( \mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1^1}, \{\mu_{\mathcal{M}_\ell^1}\}_{\ell=2}^m) + \sum_{\ell=2}^{m_1} h_{\mathcal{A}^r,\mathcal{M}_\ell^r} \right) \prod_{p\neq 1} \mathcal{A}^r(\{\mu_{\mathcal{M}_\ell^p}\}_{\ell=1}^{m_1}).$$

□

**Lemma A.9 (Specification of ComponentUIF($\mathcal{A}^j, \mathcal{M}_\ell^j$)).** *The output of* ComponentUIF($\mathcal{A}^j, \mathcal{M}_\ell^j$) *is given as*

$$\text{ComponentUIF}(\mathcal{A}^j, \mathcal{M}_\ell^j) = \sum_{\mathbf{w}_\ell^j} \mathcal{B}_\ell^j(\{\mu_{\mathcal{M}_r^j}\}_{r=1}^{m_j})\{\mathcal{V}_{\mathcal{M}_\ell^j} - \mu_{\mathcal{M}_\ell^j}\},$$

*where $\mathbf{W}_\ell^j$ is some subset of variables $\mathbf{V}$ and $\mathcal{B}_\ell^j$ is an arithmetic operator specified by running the procedure* ComponentUIF($\mathcal{A}^j, \mathcal{M}_\ell^j$).

*Proof.* Running line 1 of ComponentUIF($\mathcal{A}^j, \mathcal{M}_\ell^j$) results in $\sum_{\mathbf{w}_\ell^j} \mathcal{B}_\ell^j(\{\mathcal{M}_r^j\}_{r=1}^{m_j})\phi_{\mathcal{M}_\ell^j}$, where $\phi_{\mathcal{M}_\ell^j}$ is an IF of $\mathcal{M}_\ell^j$ equipped with a true nuisance $\eta$. Note that $\phi_{\mathcal{M}_\ell^j} = \mathcal{V}_{\mathcal{M}_\ell^j} - \mu_{\mathcal{M}_\ell^j}$ by the definition of the UIF, and the fact that $\mathcal{M}_\ell^j = \mu_{\mathcal{M}_\ell^j} \equiv \mathbb{E}_{\mathcal{V}_{\mathcal{M}_\ell^j}(\mathbf{V};\eta)}[P]$ when $\eta$ is a true nuisance. □

**Corollary A.1 (An Influence Function of $P_{\mathbf{x}}(\mathbf{y})$).** *Let the target functional $\psi \equiv P_{\mathbf{x}}(\mathbf{y})$ be given by Eq. (4). An influence function of $\psi \equiv P_{\mathbf{x}}(\mathbf{y})$, denoted $\phi_{P_{\mathbf{x}}(\mathbf{y})}$ is given as*

$$\phi_{P_{\mathbf{x}}(\mathbf{y})} = \sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j=1}^{k_d} \sum_{\ell=1}^{m_j} \text{ComponentUIF}(\mathcal{A}^j, \mathcal{M}_\ell^j) \prod_{\substack{p\neq j \\ p=1}}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}), \tag{A.34}$$

*where* $\mathcal{V}_{\mathcal{M}_\ell^j} = \mathcal{V}_{\mathcal{M}_\ell^j}(\mathbf{V};\eta)$ *is an UIF of an mSBD adjustment* $\mathcal{M}_\ell^j$ *and* $\mu_{\mathcal{M}_\ell^j} \equiv \mathbb{E}_{\mathcal{V}_{\mathcal{M}_\ell^j}}[P]$, *and*

$$\textsc{ComponentUIF}(\mathcal{A}^j, \mathcal{M}_\ell^j) = \sum_{\mathbf{w}_\ell^j} \mathcal{B}_\ell^j(\{\mu_{\mathcal{M}_r^j}\}_{r=1}^{m_j})\{\mathcal{V}_{\mathcal{M}_\ell^j} - \mu_{\mathcal{M}_\ell^j}\},$$

*where* $\mathbf{W}_\ell^j$ *is some subset of variables* $\mathbf{V}$ *and* $\mathcal{B}_\ell^j$ *is an arithmetic operator specified by running the procedure* $\textsc{ComponentUIF}(\mathcal{A}^j, \mathcal{M}_\ell^j)$.

*Proof.*

$$\begin{aligned}
\frac{\partial}{\partial t}\Psi(P_t)|_{t=0} &= \frac{\partial}{\partial t}\sum_{\mathbf{d}\setminus\mathbf{y}}\prod_{j=1}^{k_d}\mathcal{A}^j(\{\mathcal{M}_\ell^j(P_t)\}_{\ell=1}^{m_j})|_{t=0} \\
&= \sum_{\mathbf{d}\setminus\mathbf{y}}\sum_{j=1}^{k_d}\frac{\partial}{\partial t}\mathcal{A}^j(\{\mathcal{M}_\ell^j(P_t)\}_{\ell=1}^{m_j})\Big|_{t=0}\prod_{\substack{p\neq j\\p=1}}^{k_d}\mathcal{A}^p(\{\mathcal{M}_r^p\}_{r=1}^{m_p}) \\
&\stackrel{1}{=} \sum_{\mathbf{d}\setminus\mathbf{y}}\sum_{j=1}^{k_d}\sum_{\ell=1}^{m_j}\frac{\partial}{\partial t}(\mathcal{A}^j\circ\mathcal{M}_\ell^j)(P_t)\Big|_{t=0}\prod_{\substack{p\neq j\\p=1}}^{k_d}\mathcal{A}^p(\{\mathcal{M}_r^p\}_{r=1}^{m_p}) \\
&\stackrel{2}{=} \mathbb{E}_P\left[\left(\sum_{\mathbf{d}\setminus\mathbf{y}}\sum_{j=1}^{k_d}\sum_{\ell=1}^{m_j}\textsc{ComponentUIF}(\mathcal{A}^j,\mathcal{M}_\ell^j)\prod_{\substack{p\neq j\\p=1}}^{k_d}\mathcal{A}^p(\{\mathcal{M}_r^p\}_{r=1}^{m_p})\right)\cdot S(\mathbf{V})\right] \\
&\stackrel{3}{=} \mathbb{E}_P\left[\left(\sum_{\mathbf{d}\setminus\mathbf{y}}\sum_{j=1}^{k_d}\sum_{\ell=1}^{m_j}\left(\sum_{\mathbf{w}_\ell^j}\mathcal{B}_\ell^j(\{\mathcal{M}_r^j\}_{r=1}^{m_j})\phi_{\mathcal{M}_\ell^j}\right)\prod_{\substack{p\neq j\\p=1}}^{k_d}\mathcal{A}^p(\{\mathcal{M}_r^p\}_{r=1}^{m_p})\right)\cdot S(\mathbf{V})\right] \\
&\stackrel{4}{=} \mathbb{E}_P\left[\left(\sum_{\mathbf{d}\setminus\mathbf{y}}\sum_{j=1}^{k_d}\sum_{\ell=1}^{m_j}\left(\sum_{\mathbf{w}_\ell^j}\mathcal{B}_\ell^j(\{\mu_{\mathcal{M}_r^j}\}_{r=1}^{m_j})(\mathcal{V}_{\mathcal{M}_\ell^j}-\mu_{\mathcal{M}_\ell^j})\right)\prod_{\substack{p\neq j\\p=1}}^{k_d}\mathcal{A}^p(\{\mathcal{M}_r^p\}_{r=1}^{m_p})\right)\cdot S(\mathbf{V})\right],
\end{aligned}$$

where $S(\mathbf{V})$ is a score function of the parametric submodel $P_t$, and

- $\stackrel{1}{=}$ holds by the chain rule.
- $\stackrel{2}{=}$ holds since $h_{\mathcal{A}^j,\mathcal{M}_\ell^j} = \textsc{ComponentUIF}(\mathcal{A}^j,\mathcal{M}_\ell^j)$ computes $\frac{\partial}{\partial\mathcal{M}_\ell^j}(\mathcal{A}^j\circ\mathcal{M}_\ell^j)$.
- $\stackrel{3}{=}$ holds by Lemma A.9, and
- $\stackrel{4}{=}$ holds since $\phi_{\mathcal{M}_\ell^j} = \mathcal{V}_{\mathcal{M}_\ell^j} - \mu_{\mathcal{M}_\ell^j}$ and $\mu_{\mathcal{M}_\ell^j} = \mathcal{M}_\ell^j$ when $\eta$ is a true nuisance.

□

**Corollary A.2 (An Uncentered Influence Function of $P_{\mathbf{x}}(\mathbf{y})$).** *An uncentered influence function (UIF) of* $P_{\mathbf{x}}(\mathbf{y})$ *is*

$$\begin{aligned}
\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})} &= \sum_{\mathbf{d}\setminus\mathbf{y}}\mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1^1},\{\mu_{\mathcal{M}_r^1}\}_{r=2}^{m_1})\prod_{p=2}^{k_d}\mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) \\
&\quad + \sum_{(j,\ell)\neq(1,1)}^{(k_d,m_j)}\textsc{ComponentUIF}(\mathcal{A}^j,\mathcal{M}_\ell^j)\prod_{\substack{p\neq j\\p=1}}^{k_d}\mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}),
\end{aligned} \tag{A.35}$$

*where*

$$\textsc{ComponentUIF}(\mathcal{A}^j, \mathcal{M}_\ell^j) = \sum_{\mathbf{w}_\ell^j} \mathcal{B}_\ell^j(\{\mu_{\mathcal{M}_r^j}\}_{r=1}^{m_j})\{\mathcal{V}_{\mathcal{M}_\ell^j} - \mu_{\mathcal{M}_\ell^j}\},$$

*where* $\mathbf{W}_\ell^j$ *is some subset of variables* $\mathbf{V}$ *and* $\mathcal{B}_\ell^j$ *is an arithmetic operator specified by running the procedure* $\textsc{ComponentUIF}(\mathcal{A}^j, \mathcal{M}_\ell^j)$.

*Proof.* For brevity, let

$$\mathcal{C}_\ell^j \equiv \textsc{ComponentUIF}(\mathcal{A}^j, \mathcal{M}_\ell^j) \prod_{\substack{p\neq j\\ p=1}}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}).$$

Then, an influence function of $P_{\mathbf{x}}(\mathbf{y})$ in Eq. (A.34) can be rewritten as

$$\sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j=1}^{k_d} \sum_{\ell=1}^{m_j} \textsc{ComponentUIF}(\mathcal{A}^j, \mathcal{M}_\ell^j) \prod_{\substack{p\neq j\\ p=1}}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) = \sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j=1}^{k_d} \sum_{\ell=1}^{m_j} \mathcal{C}_\ell^j.$$

Then,

$$\sum_{\mathbf{d}\setminus\mathbf{y}} \sum_{j=1}^{k_d} \sum_{\ell=1}^{m_j} \mathcal{C}_\ell^j = \sum_{\mathbf{d}\setminus\mathbf{y}} \mathcal{C}_1^1 + \sum_{(j,\ell)\neq(1,1)}^{(k_d,m_j)} \mathcal{C}_\ell^j,$$

where

$$\begin{aligned}
\sum_{\mathbf{d}\setminus\mathbf{y}} \mathcal{C}_1^1 &= \sum_{\mathbf{d}\setminus\mathbf{y}} \textsc{ComponentUIF}(\mathcal{A}^1, \mathcal{M}_1^1) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) \\
&\stackrel{1}{=} \sum_{\mathbf{d}\setminus\mathbf{y}} \mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1^1}, \{\mu_{\mathcal{M}_r^1}\}_{r=2}^{m_1}) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) - \sum_{\mathbf{d}\setminus\mathbf{y}} \mathcal{A}^1(\{\mu_{\mathcal{M}_r^1}\}_{r=1}^{m_1}) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) \\
&= \sum_{\mathbf{d}\setminus\mathbf{y}} \mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1^1}, \{\mu_{\mathcal{M}_r^1}\}_{r=2}^{m_1}) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) - \sum_{\mathbf{d}\setminus\mathbf{y}} \prod_{p=1}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) \\
&\stackrel{2}{=} \sum_{\mathbf{d}\setminus\mathbf{y}} \mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1^1}, \{\mu_{\mathcal{M}_r^1}\}_{r=2}^{m_1}) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) - \Psi(P).
\end{aligned}$$

To witness $\stackrel{1}{=}$ holds, we first note that, $\mathcal{M}_1^j$ for any $j$ is a primary mSBD operator. We recall the notion of the primary mSBD operator in Lemma 4. To define, we first recap notations. Let $\mathbf{D} \equiv An(\mathbf{Y})_{G(\mathbf{V}\setminus\mathbf{X}))}$. Let $\{\mathbf{S}_i\}_{i=1}^{k_s}$ denote $C$-components of $G$. Let $\{\mathbf{D}_j\}_{j=1}^{k_d}$ denote $C$-components in $G(\mathbf{D})$. For each $\mathbf{D}_j \subseteq \mathbf{S}_i$, the primary mSBD operator for $Q[\mathbf{D}_j]$ is given as, for $j = 1, 2, \cdots, k_d$,

$$\mathcal{M}_1^j \equiv \mathcal{M}[\mathbf{a}_j | pa(\mathbf{s}_i)\setminus\mathbf{s}_i; \mathbf{s}_i\setminus\mathbf{a}_j],$$

where $\mathbf{A}_j \equiv An(\mathbf{D}_j)_{G(\mathbf{S}_i)}$. Note $Q\left[\mathbf{A}_j\right] = \mathcal{M}_1^j$. By the design of the DML-ID algorithm in Algorithm 1 (lines a.7), $Q[\mathbf{D}_j] = \mathcal{A}^j(\{\mathcal{M}_\ell^j\}_{\ell=1}^{m_j})$ is given in the form of $\sum_{\mathbf{w}_1^j} \mathcal{M}_1^j \cdot \mathcal{R}^j(\{\mathcal{M}_\ell^j\}_{\ell=2}^{m_j})$ for some arithmetic operator $\mathcal{R}^j$ and a set of variables $\mathbf{W}_1^j$ Lemma A.8. Then,

$$\mathcal{A}^1(\{\mathcal{M}_\ell^1\}_{\ell=1}^{m_1}) = \sum_{\mathbf{w}_1^1} \mathcal{M}_1^1 \mathcal{R}^1(\{\mathcal{M}_\ell^1\}_{\ell=2}^{m_1}) \tag{A.36}$$

for some function $\mathcal{R}^1$. Then,

$$\begin{aligned}
\textsc{ComponentUIF}(\mathcal{A}^1, \mathcal{M}_1^1) &= \sum_{\mathbf{w}_1^1} \mathcal{R}^1(\{\mu_{\mathcal{M}_\ell^1}\}_{\ell=2}^{m_1}) \left\{ \mathcal{V}_{\mathcal{M}_1^1} - \mu_{\mathcal{M}_1^1} \right\} && \text{(A.37)} \\
&= \sum_{\mathbf{w}_1^1} \mathcal{V}_{\mathcal{M}_1^1} \mathcal{R}^1(\{\mu_{\mathcal{M}_\ell^1}\}_{\ell=2}^{m_1}) - \sum_{\mathbf{w}_1^1} \mu_{\mathcal{M}_1^1} \mathcal{R}^1(\{\mu_{\mathcal{M}_\ell^1}\}_{\ell=2}^{m_1}) \\
&= \mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1^1}, \{\mu_{\mathcal{M}_\ell^1}\}_{\ell=2}^{m_1}) - \mathcal{A}^1(\{\mu_{\mathcal{M}_\ell^1}\}_{\ell=1}^{m_1}), && \text{(A.38)}
\end{aligned}$$

where the first equation is by the procedure $\textsc{ComponentUIF}(\mathcal{A}^1, \mathcal{M}_1^1)$ and Eq. (A.36), and the third equation holds by Eq. (A.36).

Also, $\stackrel{2}{=}$ holds by Eq. (4) and the fact that $\mu_{\mathcal{M}_\ell^j} = \mathcal{M}_\ell^j$ when $\eta$ is a true nuisance. Therefore, an UIF is given as

$$\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})} = \sum_{\mathbf{d}\setminus\mathbf{y}} \mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1^1}, \{\mu_{\mathcal{M}_r^1}\}_{r=2}^{m_1}) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) + \sum_{(j,\ell)\neq(1,1)}^{(k_d,m_j)} \mathcal{C}_\ell^j.$$

□

## Proof for Prop. 1

**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 $\psi$ given in Thm. 2 is a Neyman orthogonal score for $\psi$.*

*Proof.* Let $\eta_t \equiv \eta(P_t)$ where $P_t \equiv P(1+tg)$ for $t \in \mathbb{R}$ and $g$ a bounded mean-zero function, is a parametric submodel. For the choice of $g$, we choose $g(\mathbf{V}) = S_t(\mathbf{V}; t=0)$, a score function of the submodel $P_t$. Notice this choice satisfies the definition of the parametric submodel – a set of distribution such that the true model $P$ is included in the set ($P_0 = P$) and $P_t$ is a valid distribution (Tsiatis 2007). To see $P_t(\mathbf{v})$ is a valid distribution, consider $H(\mathbf{v}) \equiv P(\mathbf{v})(1 + S_t(\mathbf{V}; t=0))$. Note $H(\mathbf{v})$ is a valid density since $\int P(\mathbf{v}) S_t(\mathbf{v}; t=0) d\mathbf{v} = \int \frac{\partial}{\partial t} P_t(\mathbf{v}) d\mathbf{v} = \frac{\partial}{\partial t} \int P_t(\mathbf{v}) d\mathbf{v} = 0$. Then, we can view the submodel $P_t$ as a collection of distributions locating between two distributions $P$ and $H$, since $P_t = (1-t)P + tH$.

Now, we prove a given IF is a Neyman orthogonal score, following a proof of (Chernozhukov et al. 2022, Thm. 1). Consider the following:

$$
\begin{aligned}
0 &= \mathbb{E}_{P_t}\left[\phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{v}; \psi, \eta_t)\right] \\
&= \int \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{v}; \psi, \eta_t) P_t(\mathbf{v}) d\mathbf{v} \\
&= (1-t) \int \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{v}; \psi, \eta_t) P(\mathbf{v}) d\mathbf{v} + t \int \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{v}; \psi, \eta_t) H(\mathbf{v}) d\mathbf{v}.
\end{aligned}
$$

Dividing both sides by $t$, we have

$$
\frac{1}{t} \int \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{v}; \psi, \eta_t) P(\mathbf{v}) d\mathbf{v} = \int \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{v}; \psi, \eta_t) P(\mathbf{v}) d\mathbf{v} - \int \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{v}; \psi, \eta_t) H(\mathbf{v}) d\mathbf{v}.
$$

Since $\int \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{v}; \psi, \eta) P(\mathbf{v}) d\mathbf{v} = 0$, by taking $\lim_{t \to 0}$ for both sides,

$$
\begin{aligned}
&\frac{\partial}{\partial t} \int \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{v}; \psi, \eta_t) P(\mathbf{v}) d\mathbf{v} = 0 - \int \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{v}; \psi, \eta) H(\mathbf{v}) d\mathbf{v} \\
\Leftrightarrow\ &\frac{\partial}{\partial t} \mathbb{E}_P\left[\phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V}; \psi, \eta_t)\right] |_{t=0} = - \int \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{v}; \psi, \eta) H(\mathbf{v}) d\mathbf{v} = - \int \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{v}; \psi, \eta) S_t(\mathbf{v}; t=0) P(\mathbf{v}) d\mathbf{v}.
\end{aligned}
$$

That is,

$$
\begin{aligned}
\frac{\partial}{\partial t} \mathbb{E}_P\left[\phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V}; \psi, \eta_t)\right] |_{t=0} &= -\mathbb{E}_P[\phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V}; \psi, \eta) \cdot S_t(\mathbf{V}; t=0)] \\
&= -\langle \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V}), S_t(\mathbf{V}; t=0) \rangle_{\mathcal{H}},
\end{aligned}
$$

where $\mathcal{H}$ denote the Hilbert space of mean-zero measurable random functions with finite second moment, where influence functions reside, and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes its inner product (Tsiatis 2007, Chap 2,3).

Since $\phi_{P_{\mathbf{x}}(\mathbf{y})}$ is an IF of a RAL estimator (see Sec. 2), $\phi_{P_{\mathbf{x}}(\mathbf{y})}$ resides in the space orthogonal to the parametric submodel nuisance tangent space (Tsiatis 2007, Chap 4.3, Thm. 4.2). By the definition of orthogonality in Hilbert space, $\langle \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V}), S_t(\mathbf{V}; t=0) \rangle_{\mathcal{H}} = 0$. Therefore,

$$
\frac{\partial}{\partial t} \mathbb{E}_P\left[\phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V}; \psi, \eta_t)\right] |_{t=0} = 0.
$$

This implies that $\phi_{P_{\mathbf{x}}(\mathbf{y})}$ is invariant of local perturbation of $\eta_t$, implying Neyman orthogonality. This completes the proof. □

## Proof for Theorem 3

**Lemma A.10 (Simplification of the Average of the UIF).** *Let $\hat{\mathcal{V}}_{\mathcal{M}_\ell^j}$ denote the UIF $\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}$ in Eq. (6) equipped with an estimated nuisance $\hat{\eta}$; i.e., $\hat{\mathcal{V}}_{\mathcal{M}_\ell^j} = \mathcal{V}_{\mathcal{M}_\ell^j}(\mathbf{V}; \hat{\eta})$. Let $\hat{\mu}_{\mathcal{M}_\ell^j} \equiv \mathbb{E}_{\mathcal{D}}\left[\hat{\mathcal{V}}_{\mathcal{M}_\ell^j}\right]$. Then,*

$$
\mathbb{E}_{\mathcal{D}}\left[\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V}; \hat{\eta})\right] = \sum_{\mathbf{d} \setminus \mathbf{y}} \prod_{p=1}^{k_d} \mathcal{A}^p(\{\hat{\mu}_{\mathcal{M}_r^p}\}_{r=1}^{m_p}).
$$

*Proof.*

$$
\begin{aligned}
\mathbb{E}_{P'}\left[\mathcal{V}_{P_{\mathbf{x}(\mathbf{y})}}(\mathbf{V};\eta')\right] &= \mathbb{E}_{P'}\left[\sum_{\mathbf{d}\setminus\mathbf{y}} \mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1^1}, \{\mu_{\mathcal{M}_r^1}\}_{r=2}^{m_1}) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p})\right] \\
&+ \mathbb{E}_{P'}\left[\sum_{(j,\ell)\neq(1,1)}^{(k_d,m_j)} \textsc{ComponentUIF}(\mathcal{A}^j, \mathcal{M}_\ell^j) \prod_{\substack{p\neq j\\ p=1}}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p})\right] \\
&= \mathbb{E}_{P'}\left[\sum_{\mathbf{d}\setminus\mathbf{y}} \mathcal{A}^1(\mathcal{V}_{\mathcal{M}_1^1}, \{\mu_{\mathcal{M}_r^1}\}_{r=2}^{m_1}) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p})\right] \\
&+ \mathbb{E}_{P'}\left[\sum_{(j,\ell)\neq(1,1)}^{(k_d,m_j)} \left(\sum_{\mathbf{w}_\ell^j} \mathcal{B}_\ell^j(\{\mu_{\mathcal{M}_r^j}\}_{r=1}^{m_j})\{\mathcal{V}_{\mathcal{M}_\ell^j} - \mu_{\mathcal{M}_\ell^j}\}\right) \prod_{\substack{p\neq j\\ p=1}}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p})\right] \\
&\overset{1}{=} \mathbb{E}_{P'}\left[\sum_{\mathbf{d}\setminus\mathbf{y}} \left(\sum_{\mathbf{w}_1^1} \mathcal{V}_{\mathcal{M}_1^1}\mathcal{R}^1(\{\mu_{\mathcal{M}_\ell^1}\}_{\ell=2}^{m_1})\right) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p})\right] \\
&+ \mathbb{E}_{P'}\left[\sum_{(j,\ell)\neq(1,1)}^{(k_d,m_j)} \left(\sum_{\mathbf{w}_\ell^j} \mathcal{B}_\ell^j(\{\mu_{\mathcal{M}_r^j}\}_{r=1}^{m_j})\{\mathcal{V}_{\mathcal{M}_\ell^j} - \mu_{\mathcal{M}_\ell^j}\}\right) \prod_{\substack{p\neq j\\ p=1}}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p})\right] \\
&= \sum_{\mathbf{d}\setminus\mathbf{y}} \left(\sum_{\mathbf{w}_1^1} \mathbb{E}_{P'}\left[\mathcal{V}_{\mathcal{M}_1^1}\right]\mathcal{R}^1(\{\mu_{\mathcal{M}_\ell^1}\}_{\ell=2}^{m_1})\right) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) \\
&+ \sum_{(j,\ell)\neq(1,1)}^{(k_d,m_j)} \left(\sum_{\mathbf{w}_\ell^j} \mathcal{B}_\ell^j(\{\mu_{\mathcal{M}_r^j}\}_{r=1}^{m_j})\{\mathbb{E}_{P'}\left[\mathcal{V}_{\mathcal{M}_\ell^j}\right] - \mu_{\mathcal{M}_\ell^j}\}\right) \prod_{\substack{p\neq j\\ p=1}}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) \\
&= \sum_{\mathbf{d}\setminus\mathbf{y}} \left(\sum_{\mathbf{w}_1^1} \mu_{\mathcal{M}_1^1}\mathcal{R}^1(\{\mu_{\mathcal{M}_\ell^1}\}_{\ell=2}^{m_1})\right) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) \\
&+ \sum_{(j,\ell)\neq(1,1)}^{(k_d,m_j)} \left(\sum_{\mathbf{w}_\ell^j} \mathcal{B}_\ell^j(\{\mu_{\mathcal{M}_r^j}\}_{r=1}^{m_j})\cancel{\{\mu_{\mathcal{M}_\ell^j} - \mu_{\mathcal{M}_\ell^j}\}}\right) \prod_{\substack{p\neq j\\ p=1}}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) \\
&= \sum_{\mathbf{d}\setminus\mathbf{y}} \left(\sum_{\mathbf{w}_1^1} \mu_{\mathcal{M}_1^1}\mathcal{R}^1(\{\mu_{\mathcal{M}_\ell^1}\}_{\ell=2}^{m_1})\right) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) \\
&\overset{2}{=} \sum_{\mathbf{d}\setminus\mathbf{y}} \mathcal{A}^1(\{\mu_{\mathcal{M}_\ell^j}\}_{\ell=1}^{m_1}) \prod_{p=2}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}) \\
&= \sum_{\mathbf{d}\setminus\mathbf{y}} \prod_{p=1}^{k_d} \mathcal{A}^p(\{\mu_{\mathcal{M}_r^p}\}_{r=1}^{m_p}),
\end{aligned}
$$

where $\overset{1}{=}, \overset{2}{=}$ hold by Eq. (A.36). ☐

**Lemma A.11 (Doubly robustness of the UIF of the mSBD).** *Let $\mathcal{V}_{\mathcal{M}}(\mathbf{V};\{\mu_0^k,\pi_0^k\}_{k=1}^m)$ denote the UIF of the mSBD adjustment $\mathcal{M}$ given in Eq. (5). For any arbitrary nuisances $\{\mu^k,\pi^k\}_{k=1}^m$,*

$$
\mathbb{E}\left[\mathcal{V}_{\mathcal{M}}(\mathbf{V};\{\mu^k,\pi^k\}_{k=1}^m) - \mathcal{V}_{\mathcal{M}}(\mathbf{V};\{\mu_0^k,\pi_0^k\}_{k=1}^m)\right] = \sum_{k=1}^m O_P\left(\left\|\pi^k-\pi_0^k\right\|\left\|\mu^k-\mu_0^k\right\|\right). \tag{A.39}
$$

*Proof.* For $k = 1, \cdots, m$, we define a quantity $Q_k$ as follows:

$$Q_k \equiv \overline{\mu}^k(\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)}) + \sum_{r=k}^{m} \pi^{(k:r)}(\mathbf{A}^{(r-1)}, \mathbf{X}^{(r)}) I_{\mathbf{x}^{(k:r)}}(\mathbf{X}^{(k:r)})(\overline{\mu}^{r+1}(\mathbf{x}_{r+1}, \mathbf{X}^{(r)}, \mathbf{A}^{(r)}) - \mu^r(\mathbf{X}^{(r)}, \mathbf{A}^{(r-1)})),$$

where

$$\pi^{(k:r)}(\mathbf{X}^{(r)}, \mathbf{A}^{(r-1)}) \equiv \prod_{b=k}^{r} \pi^b(\mathbf{X}^{(b)}, \mathbf{A}^{(b-1)}).$$

Let $Q_{m+1} \equiv I_{\mathbf{y}}(\mathbf{Y})$. We note that $Q_1 = \mathcal{V}_{\mathcal{M}}(\mathbf{V}; \{\mu^k, \pi^k\}_{k=1}^m)$. Also, $Q_k$ can be represented recursively as a function of $Q_{k+1}$. In particular,

$$Q_{k+1} - \overline{\mu}^{k+1} = \sum_{r=k+1}^{m} \pi^{(k+1:r)}(\mathbf{A}^{(r-1)}, \mathbf{X}^{(r)}) I_{\mathbf{x}^{(k+1:r)}}(\mathbf{X}^{(k+1:r)})(\overline{\mu}^{r+1}(\mathbf{x}_{r+1}, \mathbf{X}^{(r)}, \mathbf{A}^{(r)}) - \mu^r(\mathbf{X}^{(r)}, \mathbf{A}^{(r-1)})).$$

Then, by multiplying $\pi^k I_{\mathbf{x}_k}(\mathbf{X}_k)$ on both sides, we have

$$\begin{aligned} &\pi^k I_{\mathbf{x}_k}(\mathbf{X}_k) \left( Q_{k+1} - \overline{\mu}^{k+1}(\mathbf{x}_{k+1}, \mathbf{X}^{(k)}, \mathbf{A}^{(k)}) \right) \\ &= \sum_{r=k+1}^{m} \pi^{(k:r)}(\mathbf{A}^{(r-1)}, \mathbf{X}^{(r)}) I_{\mathbf{x}^{(k:r)}}(\mathbf{X}^{(k:r)})(\overline{\mu}^{r+1}(\mathbf{x}_{r+1}, \mathbf{X}^{(r)}, \mathbf{A}^{(r)}) - \mu^r(\mathbf{X}^{(r)}, \mathbf{A}^{(r-1)})). \end{aligned}$$

Then,

$$\begin{aligned} Q_k &\equiv \overline{\mu}^k(\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)}) + \pi^k(\mathbf{A}^{(k-1)}, \mathbf{X}^{(k)}) I_{\mathbf{x}_k}(\mathbf{X}_k)(\overline{\mu}^{k+1}(\mathbf{x}_{k+1}, \mathbf{X}^{(k)}, \mathbf{A}^{(k)}) - \mu^k(\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)})) \\ &+ \sum_{r=k+1}^{m} \pi^{(k:r)}(\mathbf{A}^{(r-1)}, \mathbf{X}^{(r)}) I_{\mathbf{x}^{(k:r)}}(\mathbf{X}^{(k:r)})(\overline{\mu}^{r+1}(\mathbf{x}_{r+1}, \mathbf{X}^{(r)}, \mathbf{A}^{(r)}) - \mu^r(\mathbf{X}^{(r)}, \mathbf{A}^{(r-1)})) \\ &= \overline{\mu}^k(\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)}) + \pi^k(\mathbf{A}^{(k-1)}, \mathbf{X}^{(k)}) I_{\mathbf{x}_k}(\mathbf{X}_k)(Q_{k+1} - \mu^k(\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)})). \end{aligned}$$

We now study the relation between $Q_k$ and $\overline{\mu}_0^k(\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)})$. Specifically, we will study

$$B_k \equiv \mathbb{E}\left[ Q_k - \overline{\mu}_0^k(\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)}) | \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-2)} \right].$$

To simplify the notation, we will use $\overline{P}_k \equiv P(\cdot | \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-2)})$. With this notation, we can rewrite $B_k$ as

$$B_k \equiv \mathbb{E}_{\overline{P}_k}\left[ Q_k - \overline{\mu}_0^k(\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)}) \right].$$

Since $Q_1 = \mathcal{V}_{\mathcal{M}}(\mathbf{V}; \{\mu^k, \pi^k\})$, and $\mu_0^0 = \mathbb{E}[\overline{\mu}_0^1] = \mathcal{M}$ by Lemma A.6, it suffices to study $B_1 = \mathbb{E}[Q_1 - \overline{\mu}_0^1]$ for the error analysis. In particular, we will prove

$$B_1 = \sum_{k=1}^{m} O_P\left( \|\pi^k - \pi_0^k\| \, \|\mu^k - \mu_0^k\| \right).$$

We will prove by induction that for $j = m, \ldots, 1$,

$$B_j = \sum_{r=j}^{m} O_P\left( \|\mu_0^r - \mu^r\| \, \|\pi_0^r - \pi^r\| \right). \tag{A.40}$$

We first show that the hypothesis holds when $j = m$. To witness,

$$\begin{aligned} B_m &\equiv \mathbb{E}_{\overline{P}_m}\left[ Q_m - \overline{\mu}_0^m \right] \\ &= \mathbb{E}_{\overline{P}_m}\left[ [\overline{\mu}^m - \overline{\mu}_0^m](\mathbf{x}_m, \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-1)}) + \pi^m(\mathbf{A}^{(m-1)}, \mathbf{X}^{(m)}) I_{\mathbf{x}_m}(\mathbf{X}_m) \left\{ I_{\mathbf{y}}(\mathbf{Y}) - \mu^m(\mathbf{A}^{(m-1)}, \mathbf{X}^{(m)}) \right\} \right] \\ &\overset{1}{=} \mathbb{E}_{\overline{P}_m}\left[ [\overline{\mu}^m - \overline{\mu}_0^m](\mathbf{x}_m, \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-1)}) + \pi^m(\mathbf{A}^{(m-1)}, \mathbf{X}^{(m)}) I_{\mathbf{x}_m}(\mathbf{X}_m) [\mu_0^m - \mu^m](\mathbf{A}^{(m-1)}, \mathbf{X}^{(m)}) \right] \\ &\overset{2}{=} \mathbb{E}_{\overline{P}_m}\left[ \pi_0^m I_{\mathbf{x}_m}(\mathbf{X}_m)[\mu^m - \mu_0^m](\mathbf{X}^{(m)}, \mathbf{A}^{(m-1)}) + \pi^m I_{\mathbf{x}_m}(\mathbf{X}_m)[\mu_0^m - \mu^m](\mathbf{A}^{(m-1)}, \mathbf{X}^{(m)}) \right] \\ &= \mathbb{E}_{\overline{P}_m}\left[ \pi_0^m I_{\mathbf{x}_m}(\mathbf{X}_m)[\mu^m - \mu_0^m](\mathbf{X}^{(m)}, \mathbf{A}^{(m-1)}) - \pi^m I_{\mathbf{x}_m}(\mathbf{X}_m)[\mu^m - \mu_0^m](\mathbf{A}^{(m-1)}, \mathbf{X}^{(m)}) \right] \\ &= \mathbb{E}_{\overline{P}_m}\left[ I_{\mathbf{x}_m}(\mathbf{X}_m) \left\{ \mu^m - \mu_0^m \right\} \left\{ \pi_0^m - \pi^m \right\} (\mathbf{A}^{(m-1)}, \mathbf{X}^{(m)}) \right] \\ &= O_P\left( \|\mu_0^m - \mu^m\| \, \|\pi_0^m - \pi^m\| \right), \end{aligned}$$

where the equation $\overset{1}{=}$ holds by applying the law of total expectation to $I_{\mathbf{y}}(\mathbf{Y})$, the equation $\overset{2}{=}$ holds since, for any arbitrary function $h$,

$$\begin{aligned}
&\mathbb{E}_{\overline{P}_m}\left[\pi_0^m I_{x_m}(\mathbf{X}_m) h(\mathbf{X}_m, \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-1)})\right] \\
&\equiv \mathbb{E}_P\left[\pi_0^m I_{x_m}(\mathbf{X}_m) h(\mathbf{X}_m, \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-1)}) | \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-2)}\right] \\
&= \mathbb{E}_P\left[\frac{1}{P(\mathbf{X}_m | \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-1)})} I_{x_m}(\mathbf{X}_m) h(\mathbf{X}_m, \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-1)}) | \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-2)}\right] \\
&= \mathbb{E}_P\left[\frac{P(\mathbf{x}_m | \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-1)})}{P(\mathbf{x}_m | \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-1)})} h(\mathbf{x}_m, \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-1)}) | \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-2)}\right] \\
&= \mathbb{E}_P\left[h(\mathbf{x}_m, \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-1)}) | \mathbf{X}^{(m-1)}, \mathbf{A}^{(m-2)}\right].
\end{aligned}$$

Assume that the hypothesis holds when $j = k+1$. For $j = k$,

$$\begin{aligned}
B_k &= \mathbb{E}_{\overline{P}_k}\left[Q_k - \overline{\mu}_0^k\right] \\
&= \mathbb{E}_{\overline{P}_k}\left[\left\{\overline{\mu}^k - \overline{\mu}_0^k\right\} + \pi^k I_{\mathbf{x}_k}(\mathbf{X}_k)\left\{Q_{k+1} - \mu^k(\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)})\right\}\right] \\
&= \mathbb{E}_{\overline{P}_k}\left[\left\{\overline{\mu}^k - \overline{\mu}_0^k\right\} + \pi^k I_{\mathbf{x}_k}(\mathbf{X}_k)\left\{Q_{k+1} - \overline{\mu}_0^{k+1}\right\} + \pi^k I_{\mathbf{x}_k}(\mathbf{X}_k)\left\{\overline{\mu}_0^{k+1} - \mu^k\right\}\right] \\
&\overset{(a)}{=} \sum_{r=k+1}^{m} O_P\left(\|\mu_0^r - \mu^r\| \|\pi_0^r - \pi^r\|\right) + \mathbb{E}_{\overline{P}_k}\left[\left\{\overline{\mu}^k - \overline{\mu}_0^k\right\} + \pi^k I_{\mathbf{x}_k}(\mathbf{X}_k)\left\{\overline{\mu}_0^{k+1} - \mu^k\right\}\right],
\end{aligned}$$

where the equality $\overset{(a)}{=}$ holds since

$$\begin{aligned}
&\mathbb{E}_{\overline{P}_k}\left[\pi^k I_{\mathbf{x}_k}(\mathbf{X}_k)\left\{Q_{k+1} - \overline{\mu}_0^{k+1}\right\}\right] \\
&= \mathbb{E}\left[\pi^k(\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}) I_{\mathbf{x}_k}(\mathbf{X}_k)\left\{Q_{k+1} - \overline{\mu}_0^{k+1}(\mathbf{x}_{k+1}, \mathbf{X}^{(k)}, \mathbf{A}^{(k)})\right\} \middle| \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-2)}\right] \\
&= \mathbb{E}\left[\pi^k(\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}) I_{\mathbf{x}_k}(\mathbf{X}_k) \mathbb{E}\left[Q_{k+1} - \overline{\mu}_0^{k+1}(\mathbf{x}_{k+1}, \mathbf{X}^{(k)}, \mathbf{A}^{(k)}) \middle| \mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}\right] \middle| \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-2)}\right] \\
&= \mathbb{E}\left[\pi^k(\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}) I_{\mathbf{x}_k}(\mathbf{X}_k) \underbrace{\mathbb{E}_{\overline{P}_{k+1}}\left[Q_{k+1} - \overline{\mu}_0^{k+1}(\mathbf{x}_{k+1}, \mathbf{X}^{(k)}, \mathbf{A}^{(k)})\right]}_{=B_{k+1}} \middle| \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-2)}\right] \\
&= \sum_{r=k+1}^{m} O_P\left(\|\mu_0^r - \mu^r\| \|\pi_0^r - \pi^r\|\right),
\end{aligned}$$

where the last equality holds by the induction hypothesis. Continuing,

$$\begin{aligned}
&\mathbb{E}_{\overline{P}_k}\left[\left\{\overline{\mu}^k - \overline{\mu}_0^k\right\} + \pi^k I_{\mathbf{x}_k}(\mathbf{X}_k)\left\{\overline{\mu}_0^{k+1} - \mu^k\right\}\right] \\
&\overset{(b)}{=} \mathbb{E}_{\overline{P}_k}\left[\left\{\overline{\mu}^k - \overline{\mu}_0^k\right\} + \pi^k I_{\mathbf{x}_k}(\mathbf{X}_k)\left\{\mu_0^k - \mu^k\right\}\right] \\
&\overset{(c)}{=} \mathbb{E}_{\overline{P}_k}\left[\pi_0^k I_{\mathbf{x}_k}(\mathbf{X}_k)\left\{\mu^k - \mu_0^k\right\} + \pi^k I_{\mathbf{x}_k}(\mathbf{X}_k)\left\{\mu_0^k - \mu^k\right\}\right] \\
&= \mathbb{E}_{\overline{P}_k}\left[I_{\mathbf{x}_k}(\mathbf{X}_k)\left\{\pi_0^k - \pi^k\right\}\left\{\mu^k - \mu_0^k\right\}\right] \\
&= O_P\left(\left\|\mu_0^k - \mu^k\right\| \left\|\pi^k - \pi_0^k\right\|\right),
\end{aligned}$$

where the equality $\overset{(b)}{=}$ holds since

$$\begin{aligned}
&\mathbb{E}\left[\pi^k(\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}) I_{\mathbf{x}_k}(\mathbf{X}_k) \overline{\mu}_0^{k+1}(\mathbf{x}_{k+1}, \mathbf{X}^{(k)}, \mathbf{A}^{(k)}) \middle| \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-2)}\right] \\
&= \mathbb{E}\left[\pi^k(\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}) I_{\mathbf{x}_k}(\mathbf{X}_k) \underbrace{\mathbb{E}\left[\overline{\mu}_0^{k+1}(\mathbf{x}_{k+1}, \mathbf{X}^{(k)}, \mathbf{A}^{(k)}) \middle| \mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}\right]}_{=\mu_0^k} \middle| \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-2)}\right].
\end{aligned}$$

Also, the equality $\stackrel{(c)}{=}$ holds since

$$\begin{aligned}
&\mathbb{E}\left[\overline{\mu}^k(\mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)})\middle|\mathbf{X}^{(k-1)}, \mathbf{A}^{(k-2)}\right] \\
&\equiv \mathbb{E}\left[\mu^k(\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)})\middle|\mathbf{X}^{(k-1)}, \mathbf{A}^{(k-2)}\right] \\
&= \mathbb{E}\left[\mu^k(\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-1)})\middle|\mathbf{X}^{(k-1)}, \mathbf{A}^{(k-2)}\right] \\
&= \sum_{\mathbf{a}_{k-1}} \mu^k(\mathbf{x}_k, \mathbf{X}^{(k-1)}, \mathbf{a}_{k-1}, \mathbf{A}^{(k-2)}) P(\mathbf{a}_{k-1}|\mathbf{X}^{(k-1)}, \mathbf{A}^{(k-2)}) \\
&= \sum_{\mathbf{x}'_k, \mathbf{a}_{k-1}} \mu^k(\mathbf{x}'_k, \mathbf{X}^{(k-1)}, \mathbf{a}_{k-1}, \mathbf{A}^{(k-2)}) P(\mathbf{x}'_k, \mathbf{a}_{k-1}|\mathbf{X}^{(k-1)}, \mathbf{A}^{(k-2)}) \frac{I_{\mathbf{x}_k}(\mathbf{x}'_k)}{P(\mathbf{x}'_k|\mathbf{a}_{k-1}, \mathbf{X}^{(k-1)}, \mathbf{A}^{(k-2)})} \\
&= \sum_{\mathbf{x}'_k, \mathbf{a}_{k-1}} \mu^k(\mathbf{x}'_k, \mathbf{X}^{(k-1)}, \mathbf{a}_{k-1}, \mathbf{A}^{(k-2)}) \pi_0^k(\mathbf{x}'_k, \mathbf{X}^{(k-1)}, \mathbf{a}_{k-1}, \mathbf{A}^{(k-2)}) I_{\mathbf{x}_k}(\mathbf{x}'_k) \overline{P}_k(\mathbf{x}'_k, \mathbf{a}_{k-1}) \\
&= \mathbb{E}_{\mu^k(\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)})\pi_0^k(\mathbf{X}^{(k)}, \mathbf{A}^{(k-1)}) I_{\mathbf{x}_k}(\mathbf{X}_k)} \left[\overline{P}_k\right].
\end{aligned}$$

This shows that

$$\begin{aligned}
B_k &= \sum_{r=k+1}^{m} O_P\left(\|\mu_0^r - \mu^r\| \|\pi_0^r - \pi^r\|\right) + O_P\left(\left\|\mu_0^k - \mu^k\right\| \left\|\pi^k - \pi_0^k\right\|\right) \\
&= \sum_{r=k}^{m} O_P\left(\|\mu_0^r - \mu^r\| \|\pi_0^r - \pi^r\|\right).
\end{aligned}$$

By induction, we can conclude that, for all $k = 1, 2, \cdots, m$,

$$B_k = \sum_{r=k}^{m} O_P\left(\|\mu_0^r - \mu^r\| \|\pi_0^r - \pi^r\|\right).$$

Therefore,

$$\begin{aligned}
B_1 &= \mathbb{E}_P\left[Q_1 - \overline{\mu}_0^1(\mathbf{x}_1, \mathbf{A}_0)\right] \\
&= \sum_{r=1}^{m} O_P\left(\|\mu_0^r - \mu^r\| \|\pi_0^r - \pi^r\|\right).
\end{aligned}$$

□

**Lemma A.12 (Asymptotic Unbiasedness implies Consistency).** *Suppose an estimator $T_N$ is asymptotically unbiased to $\mu$; i.e., $\mathbb{E}_P[T_N - \mu] \to 0$ as $N \to \infty$. Suppose an estimator has vanishing variance; i.e., var$(T_N) \to 0$ as $N \to \infty$. Then, $T_N$ is a consistent estimator of $\mu$.*

*Proof.* By Markov inequality,

$$P(|T_N - \mu| > \epsilon) = P((T_N - \mu)^2 > \epsilon^2) \leq \mathbb{E}_P\left[(T_N - \mu)^2\right]/\epsilon^2.$$

Also, for $\mu_N \equiv \mathbb{E}_P[T_N]$,

$$\begin{aligned}
\mathbb{E}_P\left[(T_N - \mu)^2\right] &\leq 2\mathbb{E}_P\left[(T_N - \mu_N)^2\right] + 2(\mu_N - \mu)^2 \\
&= 2\text{var}(T_N) + 2(\mu_N - \mu)^2 \\
&\to 0.
\end{aligned}$$

where var$(T_N) + (\mu_N - \mu) \to 0$ by the given assumptions that var$(T_N) \to 0$ and $\mathbb{E}_P[T_N - \mu] = \mu_N - \mu \to 0$ as $N \to \infty$. □

**Lemma A.13 (Continuous Mapping Theorem for $L_2(P)$).** *Let $X_n, X$ denote a random sequence defined on a metric space $S$. Suppose a function $g : S \to S'$ (where $S'$ is another metric space) is continuous almost everywhere. Suppose $g$ is bounded. Then,*

$$X_n \stackrel{L_2(P)}{\to} X \implies g(X_n) \stackrel{L_2(P)}{\to} g(X).$$

*Proof.* We first note that $X_n \overset{L_2(P)}{\to} X$ implies $X_n \overset{p}{\to} X$. Then, by continuous mapping theorem, $g(X_n) \overset{p}{\to} g(X)$. Then,

$$\lim_{n\to\infty} \|g(X_n) - g(X)\|^2 = \lim_{n\to\infty} \int_{\mathcal{X}} |g(X_n) - g(X)|^2 \, d[P] \overset{*}{=} \int_{\mathcal{X}} \lim_{n\to\infty} |g(X_n) - g(X)|^2 \, d[P] = 0,$$

where the equation $\overset{*}{=}$ holds by dominated convergence theorem in $L_2(P)$ space, which is applicable since $g(X_n), g(X)$ are bounded functions (from the given condition) and $X_n \overset{p}{\to} X$. □

**Lemma A.14 (Decomposition).** *Let $f_\eta \equiv f(\mathbf{V};\eta)$ denote a finite and continuous functional and $\eta$ denote its nuisances. For some samples $\mathcal{D} \sim P$, let $T \equiv \mathbb{E}_{\mathcal{D}}[f_\eta]$. Let $\theta_0 \equiv \mathbb{E}_P[f_{\eta_0}]$ for some $\eta_0$. Let $\mathbb{E}_{\mathcal{D}-P}[f_\eta] \equiv \mathbb{E}_{\mathcal{D}}[f_\eta] - \mathbb{E}_P[f_\eta]$. Then, the following decomposition holds:*

$$\mathbb{E}_{\mathcal{D}}[f_\eta] - \theta_0 = \mathbb{E}_{\mathcal{D}-P}[f_{\eta_0}] + \mathbb{E}_{\mathcal{D}-P}[f_\eta - f_{\eta_0}] + \mathbb{E}_P[f_\eta - f_{\eta_0}]. \tag{A.41}$$

*Suppose further that*

1. *Samples used for estimating $\eta$ are independent and separate from $\mathcal{D}$; and*
2. $\|\eta - \eta_0\| = o_P(1)$.

*Then, Eq. (A.41) reduces to*

$$\mathbb{E}_{\mathcal{D}}[f_\eta] - \theta_0 = R + \mathbb{E}_P[f_\eta - f_{\eta_0}], \tag{A.42}$$

*where $R$ is a random variable converging in distribution to a zero mean normal distribution at $\sqrt{n}$ rate, where $n \equiv |\mathcal{D}|$.*

*Proof.* We first prove the equality in Eq. (A.41).

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[f_\eta] - \theta_0 &= \mathbb{E}_{\mathcal{D}}[f_\eta] - \mathbb{E}_P[f_{\eta_0}] \\ &= \mathbb{E}_{\mathcal{D}-P}[f_\eta] + \mathbb{E}_P[f_\eta - f_{\eta_0}] \\ &= \underbrace{\mathbb{E}_{\mathcal{D}-P}[f_{\eta_0}]}_{\equiv A} + \underbrace{\mathbb{E}_{\mathcal{D}-P}[f_\eta - f_{\eta_0}]}_{\equiv B} + \mathbb{E}_P[f_\eta - f_{\eta_0}].\end{aligned}$$

We now prove Eq. (A.42).

- $A$ converges in distribution to the zero-mean normal distribution at $\sqrt{N}$ rate by the central limit theorem.
- We note that a given condition $\|\eta - \eta_0\| = o_P(1)$ implies $\|f_\eta - f_{\eta_0}\| = o_P(1)$ by continuous mapping theorem for $L_2(P)$ in Lemma A.13. In particular, Lemma A.13 is applicable since $f_\eta, f_{\eta_0}$ is a bounded and continuous function, and $\|\eta - \eta_0\| = o_P(1)$. Then, $B$ converges to zero at $o_P(1/\sqrt{N})$ rate by (Kennedy et al. 2020, Lemma 2).

Then, $R \equiv A + B$ converges in distribution to the zero-mean normal distribution at $\sqrt{N}$ rate by the Slutsky's theorem. □

**Lemma A.15 (Error analysis of DML-mSBD estimator).** *The DML-mSBD estimator $\hat{\mu}_{\mathcal{M}}$ has the following property:*

1. ***Doubly Robustness**: If either $\hat{\mu}^k$ or $\hat{\pi}^k$ is correctly specified (i.e., $\hat{\mu}^k$ is a consistent estimator for $\mu_0^k$ or $\hat{\pi}^k$ is a consistent estimator for $\pi_0^k$) for $k = 1, 2, \cdots, m$, then $\hat{\mu}_{\mathcal{M}}$ is a consistent estimator for $\mathcal{M}$.*
2. ***Debiasedness**: Suppose $\|\hat{\mu}^k - \mu_0^k\| = o_P(1)$ and $\|\hat{\pi}^k - \pi_0^k\| = o_P(1)$ for all $k = 1, 2, \cdots, m$. Then, the error between the DML-mSBD estimator $\hat{\mu}_{\mathcal{M}}$ and the corresponding mSBD adjustment $\mathcal{M}$ is*

$$\hat{\mu}_{\mathcal{M}} - \mathcal{M} = R + \sum_{k=1}^{m} O_P\left(\|\hat{\pi}^k - \pi_0^k\| \|\hat{\mu}^k - \mu_0^k\|\right), \tag{A.43}$$

*where $R$ is a random variable converging to a zero mean normal distribution at $\sqrt{N}$ rate.*

*Proof.* We first show that $\hat{\mu}_{\mathcal{M}}$ is an unbiased estimator of $\mathcal{M}$:

$$\begin{aligned}\mathbb{E}[\hat{\mu}_{\mathcal{M}}] - \mathcal{M} &\overset{1}{=} \mathbb{E}\left[\mathbb{E}_{\mathcal{V}(\mathbf{V};\{\hat{\mu}^k,\hat{\pi}^k\})}[\mathcal{D}]\right] - \mathcal{M} \\ &= \mathbb{E}\left[\mathbb{E}_{\mathcal{V}(\mathbf{V};\{\hat{\mu}^k,\hat{\pi}^k\})}[\mathcal{D}]\right] - \mathbb{E}\left[\mathcal{V}(\mathbf{V};\{\mu_0^k,\pi_0^k\})\right] \\ &= \mathbb{E}\left[\mathcal{V}(\mathbf{V};\{\hat{\mu}^k,\hat{\pi}^k\})\right] - \mathbb{E}\left[\mathcal{V}(\mathbf{V};\{\mu_0^k,\pi_0^k\})\right] \\ &= \sum_{k=1}^{m} O_P\left(\|\pi^k - \pi_0^k\| \|\hat{\mu}^k - \mu_0^k\|\right) \\ &= 0,\end{aligned}$$

where $\stackrel{1}{=}$ holds by the definition of the estimator, the second equality holds since $\mathbb{E}\left[\mathcal{V}(\mathbf{V};\{\mu_0^k,\pi_0^k\})\right] = \mathcal{M}$ as shown in Lemma 3, the third equality holds by the setting where all samples are drawn from the same distribution, the fourth equality is by Lemma A.11, and the last equality holds by the given condition for the doubly robustness. Also, under the assumption that nuisances $\hat{\mu}^k$ is finite and $\hat{\pi}^k$ are strictly positive,

$$\text{var}_P(\hat{\mu}_{\mathcal{M}}) = \frac{1}{N}\text{var}_P(\mathcal{V}_{\mathcal{M}}(\mathbf{V};\hat{\eta})) \to 0,$$

as $N \to \infty$ since $\mathcal{V}_{\mathcal{M}}(\mathbf{V};\hat{\eta})$ is bounded. Therefore, by Lemma A.12, $T_N$ is a consistent estimator of $\mathcal{M}$.

We now show the debiasedness. By applying Lemmas (A.11, A.14),

$$\begin{aligned}\hat{\mu}_{\mathcal{M}} - \mathcal{M} &= R + \mathbb{E}\left[\mathcal{V}(\mathbf{V};\{\hat{\mu}^k,\hat{\pi}^k\})\right] - \mathbb{E}\left[\mathcal{V}(\mathbf{V};\{\mu_0^k,\pi_0^k\})\right] \\ &= R + \sum_{k=1}^{m} O_P\left(\left\|\pi^k - \pi_0^k\right\|\left\|\hat{\mu}^k - \mu_0^k\right\|\right).\end{aligned}$$

□

**Definition 3 (DML-ID Estimator).** Let $\mathcal{D} = \{\mathbf{V}_{(i)}\}_{i=1}^N$ denote samples drawn from $P(\mathbf{v})$. Let $\{\mathcal{D}_0,\mathcal{D}_1\}$ denote randomly split two halves of $\mathcal{D}$. Then, the DML-ID (Double Machine Learning estimator for any IDentifiable effect) $T_N$ for $\psi = P_{\mathbf{x}}(\mathbf{y})$ is constructed as follows:

1. For all $j = 1,2,\cdots,k_d$, $\ell = 1,2,\cdots,m_j$, estimate $\{\mu_0^{j,\ell,a},\pi_0^{j,\ell,a}\}_{a=1}^{r_{j,\ell}}$ as $\{\hat{\mu}^{j,\ell,a},\hat{\pi}^{j,\ell,a}\}_{a=1}^{r_{j,\ell}}$ from $\mathcal{D}_1$ where $\{\mu_0^{j,\ell,a},\pi_0^{j,\ell,a}\}_{a=1}^{r_{j,\ell}}$ are nuisances of the UIF of mSBD operator $\mathcal{M}_\ell^j$. Evaluate $\hat{\mu}_{\mathcal{M}_\ell^j} \equiv \mathbb{E}_{\mathcal{D}_0}\left[\mathcal{V}_{\mathcal{M}_\ell^j}(\mathbf{V};\{\hat{\mu}^{j,\ell,a},\hat{\pi}^{j,\ell,a}\}_{a=1}^{r_{j,\ell}})\right]$ using $\mathcal{D}_0$.
2. Let $T_N(\mathcal{D}_0;\mathcal{D}_1) \equiv \sum_{\mathbf{d}\setminus\mathbf{y}}\prod_{j=1}^{k_d}\mathcal{A}^j(\{\hat{\mu}_{\mathcal{M}_\ell^j}\}_{\ell=1}^{m_j})$.
3. Repeat steps (1-2) after switching $\mathcal{D}_0,\mathcal{D}_1$, and derive $T_N(\mathcal{D}_1;\mathcal{D}_0)$. Then,

$$T_N = \frac{T_N(\mathcal{D}_0;\mathcal{D}_1) + T_N(\mathcal{D}_1;\mathcal{D}_0)}{2}.$$

**Theorem 3 (Properties of DML-ID).** *Let $P_{\mathbf{x}}(\mathbf{y})$ be any identifiable causal effects. Let $\{\mathcal{M}_\ell^j\}_{j\in\{1,2,\cdots,k_d\},\ell\in\{1,2,\cdots,m_j\}}$ denote the mSBD adjustments that compose the expression Eq. (4). Let $\{\mu_0^{j,\ell,a},\pi_0^{j,\ell,a}\}_{a=1}^{r_{j,\ell}}$ denote the set of nuisances constituting the UIF of $\mathcal{M}_\ell^j$ given in Lemma 3, and let $\{\hat{\mu}^{j,\ell,a},\hat{\pi}^{j,\ell,a}\}_{a=1}^{r_{j,\ell}}$ denote their estimates. Assume that $\hat{\mu}^{j,\ell,a}$ is bounded and $\hat{\pi}^{j,\ell,a}$ is strictly positive and bounded for all $j,\ell,a$. Let $T_N$ be the DML-ID estimator of $P_{\mathbf{x}}(\mathbf{y})$ defined in Def. 4. Then,*

1. ***Debiasedness**:Suppose $\left\|\hat{\mu}^{j,\ell,a} - \mu_0^{j,\ell,a}\right\| = o_P(1)$ and $\left\|\hat{\pi}^{j,\ell,a} - \pi_0^{j,\ell,a}\right\| = o_P(1)$ for all $j,\ell,a$. Then,*

$$\begin{aligned}&T_N - P_{\mathbf{x}}(\mathbf{y}) \\ &= R + O_P\left(\sum_{j=1}^{k_d}\sum_{\ell=1}^{m_j}\sum_{a=1}^{r_{j,\ell}}\left\|\hat{\mu}^{j,\ell,a} - \mu_0^{j,\ell,a}\right\|\left\|\hat{\pi}^{j,\ell,a} - \pi_0^{j,\ell,a}\right\|\right),\end{aligned} \tag{A.44}$$

*where $R$ is a variable that converges to a zero-mean normal distribution* $\text{NORMAL}(0,\phi^2_{P_{\mathbf{x}}(\mathbf{y})})$ *at $\sqrt{N}$ rate, where $\phi_{P_{\mathbf{x}}(\mathbf{y})} = \phi_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V};\eta)$ is the IF of $P_{\mathbf{x}}(\mathbf{y})$ equipped with a true nuisance $\eta$ given in Thm. 2.*

2. ***Doubly Robustness**: If, $\forall j,\ell,a$, either $\hat{\mu}^{j,\ell,a}$ or $\hat{\pi}^{j,\ell,a}$ is correctly specified (i.e., $\hat{\mu}^{j,\ell,a}$ is a consistent estimator for $\mu_0^{j,\ell,a}$ or $\hat{\pi}^{j,\ell,a}$ is a consistent estimator for $\pi_0^{j,\ell,a}$), then $T_N$ is a consistent estimator for $P_{\mathbf{x}}(\mathbf{y})$.*

*Proof.* Without loss of generality, we will prove for $T_N = T_N(\mathcal{D}_0;\mathcal{D}_1)$, and set $\mathcal{D} = \mathcal{D}_0$. In the proof, we use $\mathcal{A}$ to denote the following arithmetic operator

$$\mathcal{A}(\{\mu_{\mathcal{M}_\ell^j}\}_{j,\ell}) \equiv \sum_{\mathbf{d}\setminus\mathbf{y}}\prod_{j=1}^{k_d}\mathcal{A}^j(\{\mu_{\mathcal{M}_\ell^j}\}_{\ell=1}^{m_j}).$$

Then,

$$\begin{aligned}T_N &= \mathcal{A}(\{\hat{\mu}_{\mathcal{M}_\ell^j}\}_{j,\ell}) \\ P_{\mathbf{x}}(\mathbf{y}) &= \mathcal{A}(\{\mu_{\mathcal{M}_\ell^j}\}_{j,\ell}),\end{aligned}$$

where $\mu_{\mathcal{M}_\ell^j} \equiv \mathbb{E}_P\left[\mathcal{V}_{\mathcal{M}_\ell^j}(\mathbf{V};\eta)\right]$ for the true nuisance $\eta$, and $\hat{\mu}_{\mathcal{M}_\ell^j} \equiv \mathbb{E}_{\mathcal{D}_0}\left[\mathcal{V}_{\mathcal{M}_\ell^j}(\mathbf{V};\hat{\eta})\right]$ where $\hat{\eta}$ is an estimated nuisance from $\mathcal{D}_1$ by Lemma A.10.

We first show the doubly robustness – $T_N$ is a consistent estimator for $P_{\mathbf{x}}(\mathbf{y})$. It suffices to show that each $\hat{\mu}_{\mathcal{M}_\ell^j}$ is a consistent estimator for $\mu_{\mathcal{M}_\ell^j}$, because, by continuous mapping theorem, $\mathcal{A}(\{\hat{\mu}_{\mathcal{M}_\ell^j}\})$ is a consistent estimator for $\mathcal{A}(\{\mu_{\mathcal{M}_\ell^j}\})$ when $\hat{\mu}_{\mathcal{M}_\ell^j}$ is a consistent estimator for $\mu_{\mathcal{M}_\ell^j}$ and $\mathcal{A}$ is a continuous function. Since $\mathcal{A}$ is a differentiable mapping under the condition that $\mu_{\mathcal{M}_\ell^j}$ and $\hat{\mu}_{\mathcal{M}_\ell^j}$ is strictly positive and bounded, it suffices to show that each $\hat{\mu}_{\mathcal{M}_\ell^j}$ is a consistent estimator for $\mu_{\mathcal{M}_\ell^j}$. By doubly robustness property of $\hat{\mu}_{\mathcal{M}_\ell^j}$ stated in Lemma. A.15, $\hat{\mu}_{\mathcal{M}_\ell^j}$ is a consistent estimator of $\mu_{\mathcal{M}_\ell^j}$ under given conditions. Therefore, $T_N$ is a consistent estimator of $P_{\mathbf{x}}(\mathbf{y})$.

Now we prove the debiasedness property. For $\{(a,b) : \mathcal{M}_b^a \in \{\mathcal{M}_\ell^j\}_{j,\ell}\}$, we note that $\frac{\partial}{\partial \mu_{\mathcal{M}_b^a}}\mathcal{A}(\{\mu_{\mathcal{M}_\ell^j}\}_{j,\ell})$ is given in a form of $\frac{\partial}{\partial \mu_{\mathcal{M}_b^a}}\mathcal{A}(\{\mu_{\mathcal{M}_\ell^j}\}_{j,\ell}) = \sum_{\mathbf{w}_b^a} D_b^a(\{\mu_{\mathcal{M}_\ell^j}\}_{j,\ell})$ where $\mathbf{W}_b^a$ denotes a set of variables which could possibly be an empty set and $D_b^a$ is some function.

Let $R_\ell^j \equiv \mathbb{E}_{\mathcal{D}-P}\left[\phi_{\mathcal{M}_\ell^j}\right]$ for all $j,\ell$. Then,

$$
\begin{aligned}
&T_N - P_{\mathbf{x}}(\mathbf{y}) && \text{(A.45)}\\
&= \mathcal{A}(\{\hat{\mu}_{\mathcal{M}_\ell^j}\}_{j,\ell}) - \mathcal{A}(\{\mu_{\mathcal{M}_\ell^j}\}_{j,\ell}) && \text{(A.46)}\\
&\stackrel{1}{=} \sum_{(a,b):\mathcal{M}_b^a \in \{\mathcal{M}_\ell^j\}_{j,\ell}} \sum_{\mathbf{w}_b^a} \left( D_b^a(\{\mu_{\mathcal{M}_\ell^j}\}_{j,\ell}) \left\{\hat{\mu}_{\mathcal{M}_b^a} - \mu_{\mathcal{M}_b^a}\right\} + o_P(\{\hat{\mu}_{\mathcal{M}_b^a} - \mu_{\mathcal{M}_b^a}\}) \right) && \text{(A.47)}\\
&\stackrel{2}{=} \sum_{a,b}\sum_{\mathbf{w}_b^a} \left( D_b^a(\{\mu_{\mathcal{M}_\ell^j}\}_{j,\ell}) \left\{ R_b^a + o_P(1/\sqrt{N}) + \sum_{k=1}^{r_{a,b}} O_P\left( \left\| \hat{\mu}^{a,b,k} - \mu_0^{a,b,k} \right\| \left\| \hat{\pi}^{a,b,k} - \pi_0^{a,b,k} \right\| \right) \right\} \right)\\
&\quad + \sum_{a,b}\sum_{\mathbf{w}_b^a} \left( o_P(R_b^a) + o_P(1/\sqrt{N}) + \sum_{k=1}^{r_{a,b}} O_P\left( \left\| \hat{\mu}^{a,b,k} - \mu_0^{a,b,k} \right\| \left\| \hat{\pi}^{a,b,k} - \pi_0^{a,b,k} \right\| \right) \right) && \text{(A.48)}\\
&\stackrel{3}{=} o_P(1/\sqrt{N}) + \sum_{a,b}\sum_{\mathbf{w}_b^a} D_b^a(\{\mu_{\mathcal{M}_\ell^j}\}_{j,\ell}) R_b^a + \sum_{a,b}\sum_{k=1}^{r_{a,b}} O_P\left( \left\| \hat{\mu}^{a,b,k} - \mu_0^{a,b,k} \right\| \left\| \hat{\pi}^{a,b,k} - \pi_0^{a,b,k} \right\| \right) && \text{(A.49)}\\
&\stackrel{4}{=} o_P(1/\sqrt{N}) + \sum_{a,b}\sum_{\mathbf{w}_b^a} D_b^a(\{\mu_{\mathcal{M}_\ell^j}\}_{j,\ell}) \mathbb{E}_{\mathcal{D}-P}\left[\phi_{\mathcal{M}_b^a}\right] + \sum_{a,b}\sum_{k=1}^{r_{a,b}} O_P\left( \left\| \hat{\mu}^{a,b,k} - \mu_0^{a,b,k} \right\| \left\| \hat{\pi}^{a,b,k} - \pi_0^{a,b,k} \right\| \right) && \text{(A.50)}\\
&\stackrel{5}{=} \underbrace{o_P(1/\sqrt{N}) + \mathbb{E}_{\mathcal{D}-P}\left[\phi_{P_{\mathbf{x}}(\mathbf{y})}\right]}_{\equiv R} + \sum_{a,b}\sum_{k=1}^{r_{a,b}} O_P\left( \left\| \hat{\mu}^{a,b,k} - \mu_0^{a,b,k} \right\| \left\| \hat{\pi}^{a,b,k} - \pi_0^{a,b,k} \right\| \right), && \text{(A.51)}
\end{aligned}
$$

where

1. $\stackrel{1}{=}$ holds by applying the Taylor Theorem up to the first order. We note that Taylor's theorem is applicable since $\mathcal{A}$ is smooth under the condition that $\mu^{j,\ell,a}, \hat{\mu}^{j,\ell,a} < \infty$ and $c < \pi^{j,\ell,a}, \hat{\pi}^{j,\ell,a} < \infty$ for some $c \in (0, 1/2)$.
2. $\stackrel{2}{=}$ holds by applying the error analysis $\hat{\mu}_{\mathcal{M}_b^a} - \mu_{\mathcal{M}_b^a}$ in Lemma. A.15.
3. $\stackrel{3}{=}$ holds because
   - $o_P(R_b^a) = o_P(1/\sqrt{N})$ since $R_b^a = O_P(1/\sqrt{N})$ because it converges at rate $\sqrt{N}$ by the central limit theorem, and therefore, $o_P(R_b^a) = o_P(1/\sqrt{N})$ (Van der Vaart 2000, Section 2.2).
   - For any sequence $a_N$ and a constant $c$, $o_P(a_N) + o_P(a_N) = o_P(a_N)$ and $c \cdot o_P(a_N) = o_P(a_N)$. Also, $O_P(a_N) + O_P(a_N) = O_P(a_N)$ and $c \cdot O_P(a_N) = O_P(a_N)$.
4. $\stackrel{4}{=}$ holds because of the definition $R_b^a \equiv \mathbb{E}_{\mathcal{D}-P}\left[\phi_{\mathcal{M}_b^a}\right]$.
5. $\stackrel{5}{=}$ holds since

$$
\sum_{a,b}\sum_{\mathbf{w}_b^a} D_b^a(\{\mu_{\mathcal{M}_\ell^j}\}_{j,\ell}) \mathbb{E}_{\mathcal{D}-P}\left[\phi_{\mathcal{M}_b^a}\right] = \mathbb{E}_{\mathcal{D}-P}\left[ \sum_{a,b}\sum_{\mathbf{w}_b^a} D_b^a(\{\mu_{\mathcal{M}_\ell^j}\}_{j,\ell}) \phi_{\mathcal{M}_b^a} \right],
$$

where the equation holds because (1) $\mu_{\mathcal{M}_\ell^j}$ are constants, and (2) by Coro. A.1 which states that an influence function of $P_{\mathbf{x}}(\mathbf{y})$ is given by applying the chain rule; specifically, $\phi_{P_{\mathbf{x}}(\mathbf{y})}$ is given a

$$\phi_{P_{\mathbf{x}}(\mathbf{y})} = \sum_{(a,b):\mathcal{M}_b^a \in \{\mathcal{M}_\ell^j\}_{j,\ell}} \textsc{ComponentUIF}(\mathcal{A}, \mathcal{M}_b^a) = \sum_{(a,b):\mathcal{M}_b^a \in \{\mathcal{M}_\ell^j\}_{j,\ell}} \sum_{\mathbf{w}_b^a} D_b^a(\{\mu_{\mathcal{M}_\ell^j}\}_{j,\ell})\phi_{\mathcal{M}_b^a}$$

where the first equation holds by Coro. A.1 and the second equation holds because $\textsc{ComponentUIF}(\mathcal{A}, \mathcal{M}_b^a)$ computes the partial derivative of $\mathcal{A}$ w.r.t. $\mathcal{M}_b^a$ on the direction of the influence function $\phi_{\mathcal{M}_b^a}$. As a result, $\textsc{ComponentUIF}(\mathcal{A}, \mathcal{M}_b^a)$ outputs linear function of $\phi_{\mathcal{M}_b^a}$ where its coefficients are given as a derivative of $\mathcal{A}$ w.r.t. $\mathcal{M}_b^a$; i.e., $D_b^a(\{\mu_{\mathcal{M}_\ell^j}\}_{j,\ell})$.

Finally, we note that $R$ converges in a zero-mean normal distribution $\textsc{normal}(0, \phi^2_{P_{\mathbf{x}}(\mathbf{y})})$ at $\sqrt{N}$ rate, because $\mathbb{E}_{\mathcal{D}-P}\left[\phi_{P_{\mathbf{x}}(\mathbf{y})}\right]$ converges in $\textsc{normal}(0, \phi^2_{P_{\mathbf{x}}(\mathbf{y})})$ by central limit theorem, and $\mathbb{E}_{\mathcal{D}-P}\left[\phi_{P_{\mathbf{x}}(\mathbf{y}}\right] + o_P(1/\sqrt{N})$ converges in $\textsc{normal}(0, \phi^2_{P_{\mathbf{x}}(\mathbf{y})})$ by Slutsky's theorem. □

## B Details in Experiments

The models in Examples 1 and 2 are constructed from a benchmark Bayesian network called '*Alarm*' (Beinlich et al. 1989), originally collected from a system used to monitor patients' conditions. Given the original 'alarm' network (denoted $G_{pop}$) and dataset (denoted $\mathcal{D}_{pop}$) [5], we derived the causal graphs $G$ in Fig. 1a (Example 1) and Fig. 1b (Example 2) and the corresponding datasets $\mathcal{D}$ ($N = 10000$ samples each) by marginalizing/conditioning over some variables. The exact details of how the models in Examples 1 and 2 are constructed are provided in Section B.2. All the variables in Fig. 1a and Fig. 1b are discrete. Their correspondence with the original 'Alarm' network and their domains are provided in Table (1,2) respectively.

| **Variables** | $W$ | $R$ | $X$ | $Y$ |
|---|---|---|---|---|
| **Name** | CCHL | HR | CO | BP |
| **Domain** (numeric) | $\{0, 1\}$ | $\{0, 1, 2\}$ | $\{0, 1, 2\}$ | $\{0, 1, 2\}$ |
| **Domain** | {Normal, High} | {Low, Normal, High} | {Low, Normal, High} | {Low, Normal, High} |

Table 1: Table for matching variables in Fig. 1a to the nodes in original 'Alarm' network.

| **Variables** | $X1$ | $Z$ | $R$ | $X2$ | $Y$ |
|---|---|---|---|---|---|
| **Name** | SHNT | VTUB | SAO2 | VLNG | CCHL |
| **Domain** (numeric) | $\{0, 1\}$ | $\{0, 1, 2, 3\}$ | $\{0, 1, 2\}$ | $\{0, 1, 2, 3\}$ | $\{0, 1\}$ |
| **Domain** | {Normal, High} | {Zero, Low, Normal, High} | {Low, Normal, High} | {Zero, Low, Normal, High} | {Normal, High} |

Table 2: Table for matching variables in Fig. 1b to the nodes in original 'Alarm' network.

The ground-truth values of the target causal effect $\mu(\mathbf{x}) \equiv P_{\mathbf{x}}(Y = 1)$ are computed using $G_{pop}$ and $\mathcal{D}_{pop}$. We computed the ground-truth by $\mu(\mathbf{x}) = \sum_{pa(\mathbf{x}\setminus\mathbf{x})} P_{pop}(y|\mathbf{x}, Pa(\mathbf{x})\setminus\mathbf{x}) P_{pop}(Pa(\mathbf{x})\setminus\mathbf{x})$ based on $G_{pop}$ (Pearl 2000, Thm. 3.2.2), where $P_{pop}$ is estimated from $\mathcal{D}_{pop}$.

### B.1 Background information – Marginalizing and Conditioning

In this section, we introduce operations corresponding to marginalizing and conditioning over variables in a given graph and its corresponding probability distribution.

Let $G_{pop} \equiv (\mathbf{V}_{pop}, \mathbf{E}_{pop})$ be composed of nodes $\mathbf{V}_{pop}$ and and edges $\mathbf{E}_{pop}$. Let $\mathcal{D}_{pop} = \{\mathbf{V}_{pop,(i)}\}_{i=1}^{N}$ a set of samples drawn from a distribution $P_{pop}(\mathbf{v}_{pop})$ compatible with $G_{pop}$.

**Marginalization** *Marginalizing the distribution* $P_{pop}$ over $\overline{\mathbf{C}} \equiv \mathbf{V}_{pop}\setminus\mathbf{C}$ for some $\mathbf{C}$ (i.e., $\sum_{\overline{\mathbf{c}}} P(\mathbf{v}_{pop})$) means to have $P_{pop}(\mathbf{c}) = \sum_{\overline{\mathbf{c}}} P(\mathbf{v}_{pop})$. The corresponding operation over the sample (*marginalizing the samples*) means to take $\mathcal{D}(\mathbf{c}) = \{\mathbf{C}_{(i)}\}_{i=1}^{N}$ by hiding columns corresponding to variables $\overline{\mathbf{C}}$ in $\mathcal{D}_{pop}$. This data set $\mathcal{D}(\mathbf{C})$ is a set of samples drawn from $P(\mathbf{c})$.

*Marginalizing the graph* consists of a series of graphical operation to derive $G[\mathbf{C}]$ compatible with $P(\mathbf{c})$. A series of marginalizing operations is given as the following: For each $Z \in \overline{\mathbf{C}}$, and a pair of nodes $(X, Y)$ adjacent to $Z$, add the corresponding edges between $(X, Y)$ according to Fig. B.3(a) (Koster et al. 2002) and then remove $Z$. The procedure yields a graph compatible with $P(\mathbf{V}_{pop}\setminus\{Z\})$. As a simple example, suppose $G_{pop} = \{X \leftarrow Z \rightarrow Y\}$, compatible with $P(x, y, z)$. Then, one can have a graph compatible with $P(x, y) = \sum_{z} P(x, y, z)$ by removing $Z$ and adding an edge $X \leftrightarrow Y$, following Fig. B.3(a) row 2, column 3.

**Conditioning** *Conditioning the distribution* $P_{pop}$ on $\mathbf{C} = \mathbf{c}$ means to have $P_{pop}(\overline{\mathbf{c}}|\mathbf{c})$. The corresponding operation to the sample (*conditioning the samples*) means to take $\mathcal{D}|_{\mathbf{c}} = \{\mathbf{V}_{pop,(i)}\}_{i:\mathbf{C}_{(i)}=\mathbf{c}}$ where $\mathbf{C}_{(i)} \subseteq \mathbf{V}_{pop,(i)}$. This data set $\mathcal{D}|_{\mathbf{c}}$ is a set of samples drawn from $P_{pop}(\overline{\mathbf{c}}|\mathbf{c})$.

*Conditioning the graph* on $\mathbf{C}$ consists of a series of graphical operation to derive $G|_{\mathbf{c}}$ compatible with $P_{pop}(\overline{\mathbf{c}}|\mathbf{c})$. A series of conditioning operations is given as the following: For each $Z \in \mathbf{C}$, and a pair of nodes $(X, Y)$ adjacent to $Z$, add the corresponding edges between $(X, Y)$ according to Fig. B.3(b) (Koster et al. 2002) and then remove $Z$. The procedure yields a graph compatible with $P(\mathbf{V}_{pop}|\{Z\})$. As a simple example, suppose $G_{pop} = \{X \leftrightarrow Z \leftrightarrow Y\}$, compatible with $P(x, y, z)$. Then, one can have a graph compatible with $P(x, y|z)$ by removing $Z$ and adding $X \leftrightarrow Y$, following Fig. B.3(b) row 3, column 3.

[5]The network and dataset are available at https://www.bnlearn.com/bnrepository/.

| | $Z \leftarrow Y$ | $Z \rightarrow Y$ | $Z \leftrightarrow Y$ | $Z - Y$ |
|---|---|---|---|---|
| $X \rightarrow Z$ | $\emptyset$ | $X \rightarrow Y$ | $\emptyset$ | $X - Y$ |
| $X \leftarrow Z$ | $X \leftarrow Y$ | $X \leftrightarrow Y$ | $X \leftrightarrow Y$ | $X \leftarrow Y$ |
| $X \leftrightarrow Z$ | $\emptyset$ | $X \leftrightarrow Y$ | $\emptyset$ | $X \leftarrow Y$ |
| $X - Z$ | $X - Y$ | $X \rightarrow Y$ | $X \rightarrow Y$ | $X - Y$ |

(a) Marginalizing $Z$.

| | $Z \leftarrow Y$ | $Z \rightarrow Y$ | $Z \leftrightarrow Y$ | $Z - Y$ |
|---|---|---|---|---|
| $X \rightarrow Z$ | $X - Y$ | $\emptyset$ | $X \rightarrow Y$ | $\emptyset$ |
| $X \leftarrow Z$ | $\emptyset$ | $\emptyset$ | $\emptyset$ | $\emptyset$ |
| $X \leftrightarrow Z$ | $X \leftarrow Y$ | $\emptyset$ | $X \leftrightarrow Y$ | $\emptyset$ |
| $X - Z$ | $\emptyset$ | $\emptyset$ | $\emptyset$ | $\emptyset$ |

(b) Conditioning $Z$.

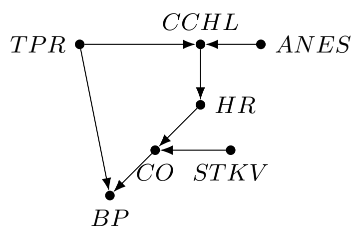
Figure B.3: An edge rendered by marginalizing and conditioning $Z = z$ (Koster et al. 2002).

**Augmentation** In an augmentation operation, we create new variables $\mathbf{C}$ using some data-generating functions $f_{\mathbf{C}}(\mathbf{W})$ for some $\mathbf{W} \subseteq \mathbf{V}_{pop}$ (i.e., $\mathbf{C} \leftarrow f_{\mathbf{C}}(\mathbf{W})$). *Augmenting variables* $\mathbf{C}$ *to the distribution* $P_{pop}$ means to have an augmented distribution $P_{pop}(\mathbf{c}, \mathbf{v}_{pop})$. The corresponding operation to the sample (*augmenting the samples*) means to take $\mathcal{D}(\mathbf{C}, \mathbf{V}_{pop}) = \{\mathbf{V}_{pop,(i)}, \mathbf{C}_{(i)}\}_{i=1}^{N}$. This data set $\mathcal{D}(\mathbf{C}, \mathbf{V}_{pop})$ is a set of samples drawn from $P_{pop}(\mathbf{c}, \mathbf{v}_{pop})$. *Augmenting the graph* means to have a graph $G = ((\mathbf{V}_{pop}, \mathbf{C}), (\mathbf{E}_{pop}, \mathbf{E}_{\mathbf{C}}))$ where $\mathbf{E}_{\mathbf{C}}$ are edges from $\mathbf{W}$ to $\mathbf{C}$.

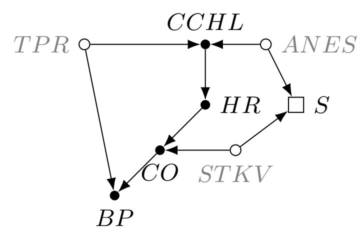
## B.2 Construction of models in Examples 1 and 2

Given the 'Alarm' network $G_{pop}$ and the data set $\mathcal{D}_{pop}$, we design a series of marginalization/conditioning/augmentation operations to reach the target graph $G$. The corresponding dataset $\mathcal{D}$ is derived accordingly as described in Section B.1.

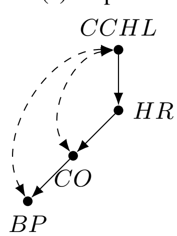
The details are described in the following.



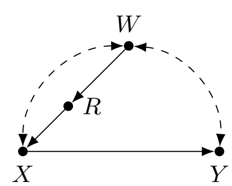

(a) Step 1




(b) Step 2




(c) Step 3




(d) Step 4

Figure B.4: The process of deriving Fig. 1a from *Alarm* network. Marginalized variables are represented in gray color. A square node (i.e., '$\square S$') is a conditioned node, where $S$ is generated by some structural causal function $S \leftarrow f_S(\text{ANES}, \text{STKV})$.

**Example 1 – Fig. 1a**

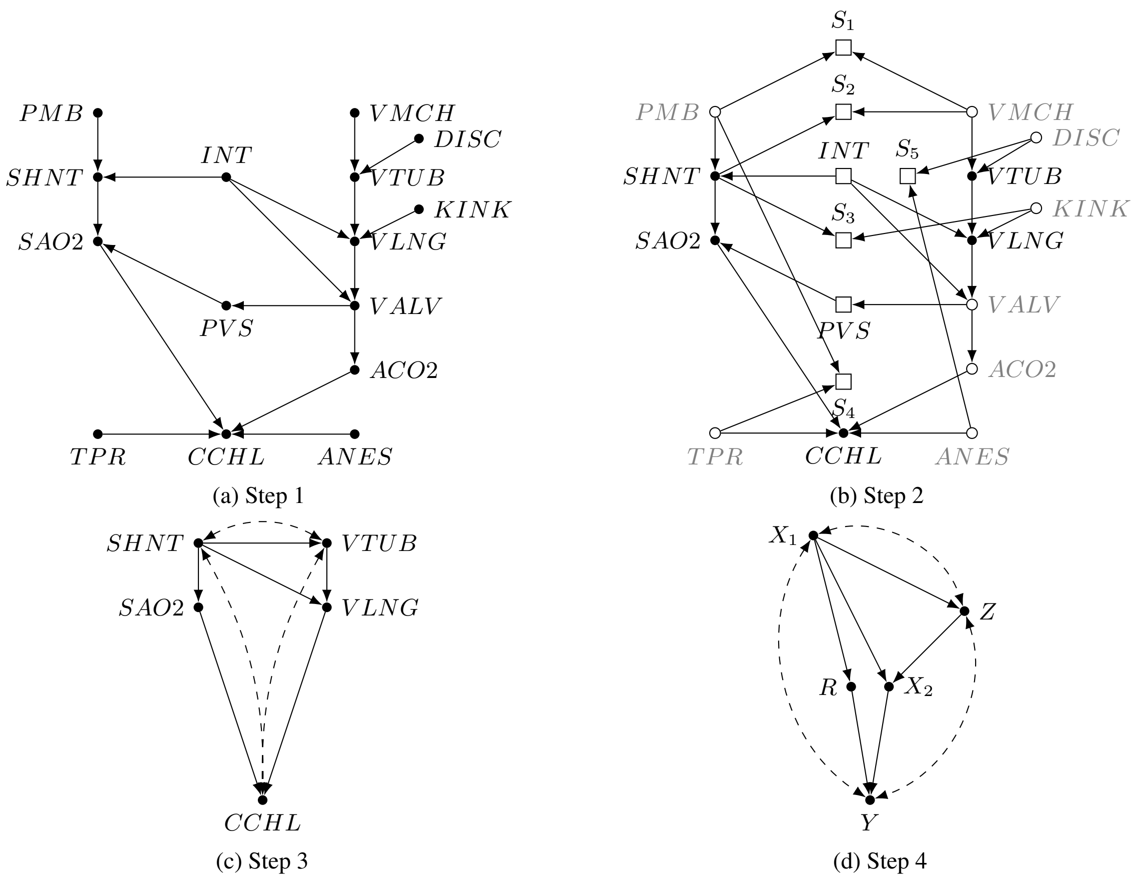
In Fig. 1a, $W = \text{CCHL}$, $R = \text{HR}$, $X = \text{CO}$, $Y = \text{BP}$.

**(Step 1)** We first take a subgraph from Alarm network that contains the set of variables of interest (marginalization). The subgraph is given as Fig. B.4a.

**(Step 2,3)** Graphs in Fig. (B.4b,B.4c) are obtained as follows:

1. We first marginalized variable TPR. By the marginalization, we will have a bidirected edge between CCHL and BP (see Fig. B.3a (row 2, column 2)), as in Fig. B.4c). The marginalized variable is marked in a gray color in Fig. B.4b.
2. Then, we augment a binary $S$ nodes using some structural causal function $S \leftarrow f_S(\text{ANES}, \text{STKV})$.
3. Then, condition on samples with $S = 1$. This procedure generates a conditioned node $\square S$ in Fig. B.4b. Notice that this conditioning generates an edge CCHL $\leftarrow$ ANES $-$ STKV $\rightarrow$ CO (see Fig. B.3b (row 1, column 1)).
4. We then marginalizing ANES, STKV, in turn. By marginalizing over ANES, we have CCHL $\leftarrow$ STKV $\rightarrow$ CO (see Table. B.3b (row 2, column 4)). By marginalizing STKV, we have CCHL $\leftrightarrow$ CO (see Fig. B.3a (row 2, column 2)), as shown in Fig. B.4c.

**(Step 4)** By setting $W$ = CCHL, $R$ = HR, $X$ = CO, $Y$ = BP and rearranging positions of nodes, we obtain the desired graph.




(a) Step 1 (b) Step 2 (c) Step 3 (d) Step 4

Figure B.5: The process of deriving Fig. 1b from *Alarm* network. Marginalized variables are represented in gray color. A square node (e.g., '$\square S_i$', for $i = 1, \cdots, 5$) is a conditioned node, where $S_i$ is generated by some structural causal function $S_i \leftarrow f_{S_i}(Pa_{S_i})$ where $Pa(S_i)$ is a parental set of $S_i$.

**Example 2 – Fig. 1b**
In Fig. 1b, $X_1$ = SHNT, $Z$ = VTUB, $R$ = SAO2, $X_2$ = VLNG, $Y$ = CCHL.

**(Step 1)** We first take a subgraph from Alarm network that contains the set of variables of interest. The subgraph is given as Fig. B.5a.

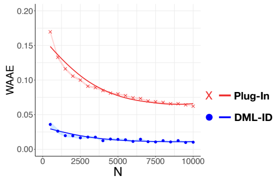
**(Step 2,3)** The graphs in Fig. (B.5b,B.5c) are obtained as follows:

1. We augment a set of binary $S_i$ nodes using some structural causal function $S_i \leftarrow f_{S_i}(\cdot)$. Specifically, $S_1 \leftarrow f_{S_1}$(PMB, VMCH), $S_2 \leftarrow f_{S_2}$(SHNT, VMCH), $S_3 \leftarrow f_{S_3}$(SHNT, KINK), $S_4 \leftarrow f_{S_4}$(PMB, TPR), $S_5 \leftarrow f_{S_5}$(VMCH, ANES).
2. Then, we conditioned on samples with $S_1 = 1, S_2 = 1, \cdots, S_5 = 1$. This procedure generates conditioned node $\square S_i$.
3. We conditioned on variables INT, PVS for blocking paths not included in a target graph. This generates $\square$INT, $\square$PVS.
4. We then marginalizing gray-colored variables in Fig. B.5b. This generates a causal graph in Fig. B.5c.

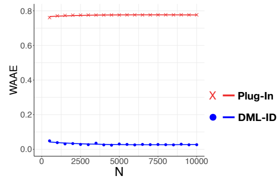
**(Step 4)** By setting $X_1$ = SHNT, $Z$ = VTUB, $R$ = SAO2, $X_2$ = VLNG, $Y$ = CCHL and rearranging positions of nodes, we obtain the desired graph.

## B.3 Additional experimental results

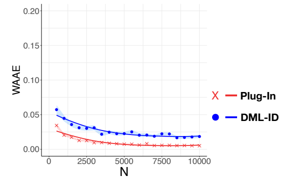
**On higher dimensional dataset.** In this section, we test the DML-ID estimator on synthetic data sets of higher dimensional. We use the causal graphs in Fig. 1a and 1b to generate synthetic data sets. For Fig. 1a, all variables are set to be binary except $W$ **is** $D$**-dimensional binary**. For Fig. 1b, all variables are set to be binary except $Z$ **is** $D$**-dimensional binary**. We performed experiments with $D = 20$.



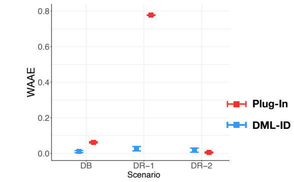

(a) Fig. 1a, 'DB'



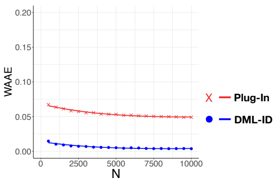

(b) Fig. 1a, 'DR-1'



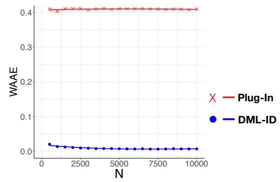

(c) Fig. 1a, 'DR-2'



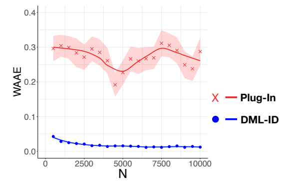

(d) Fig. 1a, Error bar



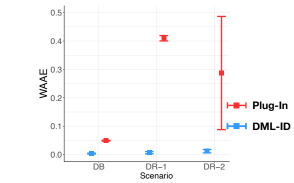

(e) Fig. 1b, 'DB'




(f) Fig. 1b, 'DR-1'

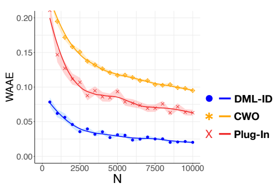



(g) Fig. 1b, 'DR-2'

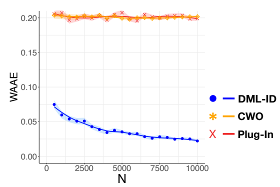



(h) Fig. 1b, Error bar

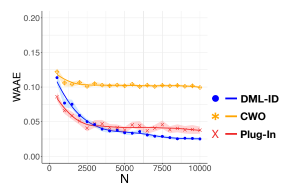
Figure B.6: Plots for **(Top)** Fig. 1a, and **(Bottom)** Fig. 1b in which $D = 20$. **(a,b,c),(e,f,g)** WAAE plots for scenarios 'Debiasedness' ('DB'), 'Doubly Robustness' ('DR-1' and 'DR-2'). **(d,h)** Error bar charts comparing WAAE at $N = 10,000$ for Fig. (1a,1b). Shades are representing standard deviation. Plots are best viewed in color.



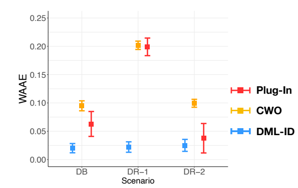

(a) Example 1, 'DB'




(b) Example 1, 'DR-1'




(c) Example 1, 'DR-2'




(d) Example 1, Error bar

Figure B.7: Plots for Example 1 comparing the proposed estimator ('DML-ID') with the plug-in and CWO estimator. **(a,b,c)** WAAE plots for scenarios 'Debiasedness' ('DB'), 'Doubly Robustness' ('DR-1' and 'DR-2'). **(d)** Error bar charts comparing WAAE at $N = 10,000$ for Example (1,2). Shades are representing standard deviation. Plots are best viewed in color.

We specify a SCM $M$ for each causal graph and generate data sets $\mathcal{D}$ from $M$. In order to estimate the ground truth $\mu(\mathbf{x}) \equiv \mathbb{E}_{P_{\mathbf{x}}}[Y]$, we generate $m_{int} = 10^7$ samples $\mathcal{D}_{int}$ from $M_{\mathbf{x}}$, the model induced by the intervention $do(\mathbf{X} = \mathbf{x})$, and compute the mean of $Y$ in $\mathcal{D}_{int}$. The code for generating the data sets are provided at the end of this section.

**Debiasedness (DB)** The WAAE plots for the debiasedness experiments are shown in Fig. B.6 (a) and (e) for Fig. 1a and 1b, respectively. The DML-ID estimator exhibits the debiasedness property against the converging noise decaying at $N^{-1/4}$ rates, while the PI estimator converges much slower, for both Fig. (1a,1b)

**Doubly robustness (DR)** The WAAE plots for the doubly robustness experiments are shown in Fig. B.6 (b, c) for Fig. 1a and (f, g) for Fig. 1b. Two misspecification scenarios are simulated for each example. For Fig. 1a, 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, resulting in that the DML-ID estimator is compared with the correctly specified PI estimator. For Fig. 1b, nuisance $\{P(y|x_1, x_2, r, z), P(x_1, z)\}$ are misspecified in 'DR-1', and $\{P(r, x_2|x_1, z)\}$ is 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, or DR-2 for Fig. 1b).

Finally, to further assess the performance of DML-ID when compared against PI, we present the error bar chart of averages and $\pm 1$ standard deviations of WAAEs with the fixed $N = 10,000$ for each of the three scenarios (DB, DR-1, DR-2) in Fig. B.6 (d) for Fig. 1a and in Fig. B.6 (h) for Fig. 1b.

**Comparison with other estimators.** To answer the feedback of the reviewer, we compared our DML-ID estimator with the estimator ('CWO') proposed by (Jung, Tian, and Bareinboim 2020a). We note that CWO covers some special settings and are applicable to Example 1 (Fig. 1a), but not to Example 2 (Fig. 1b). The result indicates that the DML-ID estimator outperforms the CWO estimator, enjoying debiasedness and doubly robustness. This result will be incorporated into the paper.

## 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. 2022. Locally robust semiparametric estimation. *Econometrica* 90(4): 1501–1535.

Colangelo, K.; and Lee, Y.-Y. 2020. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036* .

Díaz, I.; and van der Laan, M. J. 2013. Targeted data adaptive estimation of the causal dose–response curve. *Journal of Causal Inference* 1(2): 171–192.

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* .

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.

Huang, Y.; and Valtorta, M. 2008. On the completeness of an identifiability algorithm for semi-markovian models. *Annals of Mathematics and Artificial Intelligence* 54(4): 363–408.

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 *Proc. 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. 2020a. Efficient nonparametric causal inference with missing exposure information. *The international journal of biostatistics* 16(1).

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

Kennedy, E. H. 2022. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469* .

Kennedy, E. H.; Balakrishnan, S.; G'Sell, M.; et al. 2020. Sharp instruments for classifying compliers and generalizing causal effects. *Annals of Statistics* 48(4): 2008–2030.

Kennedy, E. H.; Lorch, S.; and Small, D. S. 2019. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81(1): 121–143.

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.

Klaassen, C. A. 1987. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics* 1548–1562.

Koster, J. T.; et al. 2002. Marginalizing and conditioning in graphical models. *Bernoulli* 8(6): 817–840.

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. 2021. Doubly robust nonparametric instrumental variable estimators for survival outcomes. *Biostatistics (Oxford, England)* .

Marsden, J. E.; Hoffman, M. J.; et al. 1993. *Elementary classical analysis*. Macmillan.

Neugebauer, R.; and van der Laan, M. 2007. Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference* 137(2): 419–434.

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.

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.; Li, L.; Tchetgen, E.; van der Vaart, A.; et al. 2008. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, 335–421. Institute of Mathematical Statistics.

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

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

Robins, J. M.; and Ritov, Y. 1997. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in medicine* 16(3): 285–319.

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. 2020. Efficient Adjustment Sets for Population Average Causal Treatment Effect Estimation in Graphical Models. *Journal of Machine Learning Research* 21(188): 1–86.

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

Rudolph, K. E.; and van der Laan, M. J. 2017. Robust estimation of encouragement-design intervention effects transported across sites. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 79(5): 1509.

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. 2022. Efficient adjustment sets in causal graphical models with hidden variables. *Biometrika* 109(1): 49–65.

Stein, C.; et al. 1956. Efficient nonparametric testing and estimation. In *Proc. of the third Berkeley symposium on mathematical statistics and probability*, volume 1, 187–195.

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.

Toth, B.; and van der Laan, M. 2016. TMLE for marginal structural models based on an instrument. UC Berkeley Division of Biostatistics Working Paper Series. Technical report, working paper 350.

Tsiatis, A. 2007. *Semiparametric theory and missing data*. Springer Science & Business Media.

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.

Zheng, W.; and van der Laan, M. J. 2011. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, 459–474. Springer.