Estimating Identifiable Causal Effects on Markov Equivalence Class through Double Machine Learning

Yonghan Jung [1] **Jin Tian** [2] **Elias Bareinboim** [3]

Abstract

General methods have been developed for estimating causal effects from observational data under causal assumptions encoded in the form of a causal graph. Most of this literature assumes that the underlying causal graph is completely specified. However, only observational data is available in most practical settings, which means that one can learn at most a Markov equivalence class (MEC) of the underlying causal graph. In this paper, we study the problem of causal estimation from a MEC represented by a partial ancestral graph (PAG) - learnable by structural learning algorithms. We develop a general estimator for any identifiable causal effects in PAGs. The result fills a gap for an end-to-end solution to causal inference from observational data to effects estimation. Specifically, we develop a complete identification algorithm that derives an influence function for any identifiable causal effects in a PAG. We then construct a double/debiased machine learning (DML) estimator that is robust to model misspecification and biases in nuisance function estimation, permitting the use of modern machine learning techniques. Simulation results corroborate with the theory.

1. Introduction

Inferring causal effects from observational data is a fundamental task in machine learning and various empirical sciences. There exists a growing literature studying the conditions under which causal conclusions can be drawn from non-experimental data (Pearl, 2000; Bareinboim & Pearl, 2016; Pearl & Mackenzie, 2018). In particular, the literature of *causal effect identification* (Pearl, 2000, Def. 3.2.4) investigates whether, given a causal graph G encoding qualitative knowledge about the domain, 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, can be uniquely inferred from the observational distribution $P(V)$ (Pearl, 1995; Tian & Pearl, 2003; Huang & Valtorta, 2006; Shpitser & Pearl, 2006; Lee & Bareinboim, 2020). There is also a large literature on estimating causal effects from finite samples drawn from $P(V)$ when the corresponding causal estimand is in the form of covariate adjustment (or its sequential variants) (Rosenbaum & Rubin, 1983; Pearl & Robins, 1995; Robins et al., 2000; Bang & Robins, 2005; Van Der Laan & Rubin, 2006; Hill, 2011), including doubly robust estimators for addressing model misspecification (Robins et al., 1994; Bang & Robins, 2005; Van Der Laan & Rubin, 2006; Rotnitzky & Smucler, 2020; Smucler et al., 2020; Fulcher et al., 2020). Recently, machine learning (ML) based methods have been developed for estimating any causal effects from finite samples whenever they are identifiable given a causal graph (Jung et al., 2020a;b; 2021).

Despite the power of these results, their applicability is contingent upon one having a causal graph, which may be hard to mannually specify. In practical settings, one may attempt to learn the causal graph using structural learning algorithms from the available observational data (Pearl, 2000; Spirtes et al., 2000; Peters et al., 2017). Still, in principle, only a *Markov equivalence class (MEC)* of the underlying causal graph can be inferred from non-experimental data (Spirtes et al., 2000; Zhang, 2008b). There is a growing interest in causal identification in MECs (Zhang, 2008a; Perkovic et al., 2017; Jaber et al., 2018a;b). In particular, a complete algorithm (named IDP) has recently been developed for identifying causal effects in a MEC represented by a *partial ancestral graph (PAG)* (Jaber et al., 2019). PAGs are learnable from data using causal structural learning algorithms (e.g. FCI (Zhang, 2008b)),

Even though these are quite general results, it remains an open challenge to estimate the resulting causal expressions from finite samples. For concreteness, consider the PAG in Fig. 1 as an example. The IDP algorithm identifies $P_x(y_1, y_2, y_3, y_4) =$

[*]Equal contribution [1]Department of Computer Science, Purdue University, USA [2]Department of Computer Science, Iowa State University, USA [3]Department of Computer Science, Columbia University, USA. Correspondence to: Yonghan Jung <jung222@purdue.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, PMLR 108, 2021. Copyright 2021 by the author(s).

$P(y_4|y_3, y_2, y_1, x, r)P(y_1)\sum_r P(y_2, y_3|x, r)P(r)$. The only viable general-purpose method currently available for estimating arbitrary causal estimands like this is the "plug-in" estimators (Casella & Berger, 2002), which estimate each conditional probability in the estimand (e.g., $P(y_4|y_3, y_2, y_1, x, r)$), called *nuisance functions* or *nuisances* in short, often by assuming a parametric model, and plug them into the equation. However, plug-in estimators are vulnerable to model misspecification in that all nuisance models need to be correctly specified for the estimator to be consistent. They also often suffer from biases in estimating the nuisances. In recent years, it is common to learn nuisance functions using highly flexible ML models, particularly in high-dimensional settings, including methods such as random forests, boosted regression trees, and deep neural networks. In practice, these ML methods inherently trade off regularization bias with overfitting often causing acute bias in the plug-in estimators of the target estimand such that these estimators will not achieve desirable $\sqrt{N}$-consistency (Chernozhukov et al., 2018), where N is the sample size.

We will exploit in this paper the *double/debiased machine learning* (DML) framework proposed in (Chernozhukov et al., 2018). This framework provides estimators that achieve $\sqrt{N}$-consistency with respect to the target estimand while admitting the use of highly flexible ML methods for estimating the nuisances at a slower $N^{-1/4}$ rate convergence ('*debiasedness*'). DML has been applied in causal inference in some specific settings, including in the context of the backdoor/ignorability and instrumental variables (Zadik et al., 2018; Syrgkanis et al., 2019; Foster & Syrgkanis, 2019; Chernozhukov et al., 2019; Kallus & Uehara, 2020; Farbmacher et al., 2020). Recently, DML has been used for estimating causal effects when the causal graph is fully specified (Jung et al., 2021).

Our goal will be to develop a general estimator for any identifiable causal effects in PAGs (when the causal graph is unknown). In particular, we will develop a DML estimator for identifiable causal effects in PAGs, named *DML-IDP*, by deriving their *influence functions (IF)* based on the semiparametric theory (Van der Vaart, 2000). Our results fill in a gap for a purely data-driven, end-to-end solution to causal effects estimation, i.e., from observational data $\mathcal{D} = \{\mathbf{V}_{(i)}\}_{i=1}^N \rightarrow$ PAG G by structure learning algorithm → identifiability of target effect $P_x(y)$ by IDP → **estimating** $P_x(y)$ **from** $\mathcal{D}$ **by DML-IDP**. Specifically, our contributions are as follows:

1. We develop a complete systematic procedure that derives an IF for any identifiable causal effects in a PAG.

2. We develop a DML estimator (DML-IDP) for any identifiable causal effects in a PAG, which enjoy debiasedness and doubly robustness against model misspecification and biases in nuisances estimation. Experimental studies corroborate with the theory.

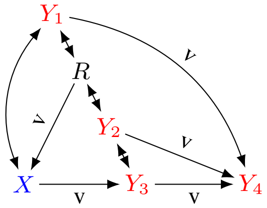

Figure 1: An example PAG. Nodes representing the treatment (X) and outcome ($\mathbf{Y}$) are marked in blue and red respectively. Causal effect $P_x(\mathbf{y})$ is identifiable.

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

2. Preliminaries

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. 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. For function $f(\mathbf{v})$ and a distribution $P(\mathbf{v})$, $\mathbb{E}_P[f(\mathbf{V})] \equiv \sum_{\mathbf{v}} f(\mathbf{v})P(\mathbf{v})$, and $\|f(\mathbf{V})\|_2 \equiv \sqrt{\mathbb{E}_P[(f(\mathbf{V}))^2]}$. $\widehat{f}$ is said to converge to f at rate r_N if $\|\widehat{f}(\mathbf{V}) - f(\mathbf{V})\|_2 = o_P(1/r_N)$.

Structural Causal Models. We use the language of structural causal models (SCMs) as our basic semantical framework (Pearl, 2000). Each SCM M over a set of variables $\mathbf{V}$ induces a distribution $P(\mathbf{v})$ and a causal graph G that is a directed acyclic graph (DAG) with birected arrows where solid-directed arrows encode functional relationships between observed variables, and bidirected arrows encode unobserved latent variables. 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 $P(\mathbf{v}|do(\mathbf{x})) \equiv P_{\mathbf{x}}(\mathbf{v})$.

Partial Ancestral Graphs (PAGs). Given non-experimental data, only a *Markov equivalence class (MEC)* of the underlying causal graph can be inferred that includes a set of graphs with the same conditional independences (Zhang, 2007). A PAG provides a graphical representation of a MEC. PAGs may contain directed ($\rightarrow$) or bidirected ($\leftrightarrow$) edges, representing ancestral relations, and edges with circles (e.g., $\{\circ\!\!\rightarrow, \circ\!\!-\!\!\circ\}$) indicating structural uncertainty (see Figs. 1 and 2 for example PAGs).

Given a PAG, a path between X and Y is *potentially directed* from X to Y if there is no arrowhead $\{<, >\}$ on the path pointing towards X. Y is called a *possible descendant* of X and X a *possible ancestor* of Y and denoted $X \in An(Y)$ if there is a potentially directed path from X to Y. Y is called a *possible child* of X and denoted $Y \in Ch(X)$, and X a *possible parent* of Y and denoted $X \in Pa(Y)$, if they are adjacent and the edge is not into X. By stipulation, $X \in An(X)$, $X \in Pa(X)$, and $X \in Ch(X)$. For a set of nodes $\mathbf{X}$, we have $Pa(\mathbf{X}) = \bigcup_{X \in \mathbf{X}} Pa(X)$ and $Ch(\mathbf{X}) = \bigcup_{X \in \mathbf{X}} Ch(X)$. If the edge marks on a path between X and Y are all circles, we call the path a circle path. We refer to the closure of nodes connected with circle paths as a *bucket*. Nodes $\mathbf{V}$ in a PAG G are partitioned into a unique set of buckets $\mathbf{V} = \bigcup_{i=1}^{n} \mathbf{B}_i$. There exists a topological order over buckets $\mathbf{B}_1 \prec \cdots \prec \mathbf{B}_n$ that defines a partial order over $\mathbf{V}$, which is valid in all the causal graphs in the MEC. This is named a *partial topological order (PTO)* and could be assigned by (Jaber et al., 2018a, Algo. 2). Given a PTO $\prec$ and a set $\mathbf{C} \subseteq \mathbf{V}$, we denote $\text{pre}_{\mathbf{C}}(\mathbf{B}_i) \equiv (\bigcup_{j<i} \mathbf{B}_j) \cap \mathbf{C}$ and use $\text{pre}(\mathbf{B}_i) \equiv \text{pre}_{\mathbf{V}}(\mathbf{B}_i)$. An *inducing path* is a path on which every node V_i (except for the endpoints) is a *collider* on the path and every collider is an ancestor of an endpoint. A directed edge $X \to Y$ in a PAGis *visible* and denoted $X \overset{v}{\to} Y$ if there exists no causal graph in the corresponding MEC where there is an inducing path between X and Y that is into X. Given a PAG G and a set $\mathbf{C} \subseteq \mathbf{V}$, $G(\mathbf{C})$ denotes the subgraph composed of nodes $\mathbf{C}$ and edges therein.

Causal Effect Identification. Given a DAG G over $\mathbf{V}$, an effect $P_{\mathbf{x}}(\mathbf{y})$ where $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ is *identifiable* if $P_{\mathbf{x}}(\mathbf{y})$ is computable from the distribution $P(\mathbf{v})$ in any SCM that induces G (Pearl, 2000, p. 77). One key notion is called *confounded components (for short, C-component)* : closures of nodes connected with a path composed solely of bi-directed edges $V_i \leftrightarrow V_j$ (Tian & Pearl, 2002).

Given a PAG G over $\mathbf{V}$, a query $P_{\mathbf{x}}(\mathbf{y})$ is *identifiable* if and only if $P_{\mathbf{x}}(\mathbf{y})$ is identifiable with the same expression in every DAG in the MEC represented by the PAG G. A complete identification algorithm in PAGs called IDP has been developed (Jaber et al., 2019) (also presented in Appendix A for convenience) based on *possible C-component* (PC-component) and *definite C-component* (DC-component):

Definition 1 (PC **&** DC**-component** (Jaber et al., 2018a)**).** In a PAG (or its subgraph), two nodes are in the same PC-*component* if there is a path between them s.t. (1) all non-endpoint nodes along the path are colliders, and (2) none of the edges is visible. Two nodes are in the same DC-*component* if they are connected with a bi-directed path.

For a set of variables $\mathbf{X}$, we will use $\mathcal{C}(\mathbf{X})$ to denote the union of the PC-components that contain variables in $\mathbf{X}$. For any $\mathbf{C} \subseteq \mathbf{V}$, the quantity $Q[\mathbf{C}] \equiv P_{\mathbf{v} \setminus \mathbf{c}}(\mathbf{c})$, called a *C-factor*, is defined as the distribution of $\mathbf{C}$ under an intervention on $\mathbf{V} \setminus \mathbf{C}$. IDP algorithm is based on the following results for identification and decomposition of C-factors.

Proposition 1 ((Jaber et al., 2018b)**).** *Let G be a PAG over $\mathbf{V}$, $\mathbf{T} = \cup_{i=1}^{m} \mathbf{B}_i$ be the union of a set of buckets, and $\mathbf{X} \subseteq \mathbf{T}$ be a bucket. Given $P_{\mathbf{v} \setminus \mathbf{t}}$ (i.e., $Q[\mathbf{T}]$) and a PTO $\mathbf{B}_1 \prec \cdots \prec \mathbf{B}_m$ with respect to $G(\mathbf{T})$, $Q[\mathbf{T} \setminus \mathbf{X}]$ is identifiable if and only if $\mathcal{C}(\mathbf{X}) \cap Ch(\mathbf{X}) \subseteq \mathbf{X}$ in $G(\mathbf{T})$. If identifiable, then $Q[\mathbf{T} \setminus \mathbf{X}] = \frac{P_{\mathbf{v} \setminus \mathbf{t}}}{\mathcal{Q}_{\mathbf{S}_{\mathbf{X}}}} \sum_{\mathbf{x}} \mathcal{Q}_{\mathbf{S}_{\mathbf{X}}}$, where $\mathcal{Q}_{\mathbf{S}_{\mathbf{X}}} \equiv \prod_{i | \mathbf{B}_i \subseteq \mathbf{S}_{\mathbf{X}}} P_{\mathbf{v} \setminus \mathbf{t}}(\mathbf{b}_i | pre_{\mathbf{T}}(\mathbf{b}_i))$ and $\mathbf{S}_{\mathbf{X}} = \bigcup_{X \in \mathbf{X}} \mathbf{S}_X$ with $\mathbf{S}_X$ being the DC-component of X in $G(\mathbf{T})$.*

Definition 2 (**Region** $\mathcal{R}_{\mathbf{A}}^{\mathbf{C}}$ (Jaber et al., 2019)**).** Given a PAG G over $\mathbf{V}$ and $\mathbf{A} \subseteq \mathbf{C} \subseteq \mathbf{V}$, the *region* of $\mathbf{A}$ w.r.t. $\mathbf{C}$, denoted $\mathcal{R}_{\mathbf{A}}^{\mathbf{C}}$, is the union of the buckets in $G(\mathbf{C})$ that contain nodes in the PC-component $\mathcal{C}(\mathbf{A})$ of $\mathbf{A}$ in $G(\mathbf{C})$.

Proposition 2 ((Jaber et al., 2019)**).** *Given a PAG G over $\mathbf{V}$ and a set $\mathbf{C} \subseteq \mathbf{V}$, $Q[\mathbf{C}]$ can be decomposed as $Q[\mathbf{C}] = \frac{Q[\mathcal{R}_{\mathbf{A}}] \cdot Q[\mathcal{R}_{\mathbf{C} \setminus \mathcal{R}_{\mathbf{A}}}]}{Q[\mathcal{R}_{\mathbf{A}} \cap \mathcal{R}_{\mathbf{C} \setminus \mathcal{R}_{\mathbf{A}}}]}$ for any $\mathbf{A} \subseteq \mathbf{C}$, where $\mathcal{R}_{(\cdot)} = \mathcal{R}_{(\cdot)}^{\mathbf{C}}$.*

Semiparametric Theory. We aim to estimate a target estimand $\psi \equiv \Psi(P)$ that is a functional of $P(\mathbf{V})$ (e.g., $\Psi(P) = \sum_z P(y|x,z)P(z)$) from finite samples $\mathcal{D} = \{\mathbf{V}_{(i)}\}_{i=1}^{N}$ drawn from P. Let a *parametric submodel* $P_t \equiv P(\mathbf{v})(1 + tg(\mathbf{v}))$ for any $t \in \mathbb{R}$ and bounded mean-zero function $g(\cdot)$ over random variables $\mathbf{V}$. If a functional $\Psi(P_t)$ is pathwise (formally, Gâteaux) differentiable at $t = 0$, then there exists a function $\phi(\mathbf{V}; \psi, \eta)$ (shortly ϕ), called an *influence function (IF)* for ψ, where $\eta = \eta(P)$ stands for the set of nuisance functions comprising ϕ, satisfying $\mathbb{E}_P[\phi] = 0$, $\mathbb{E}_P[\phi^2] < \infty$, and $\frac{\partial}{\partial t} \Psi(P_t)|_{t=0} = \mathbb{E}_P[\phi(\mathbf{V}; \psi, \eta) S_t(\mathbf{V}; t = 0)]$ 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). Given an IF ϕ, a Regular and Asymptotic Linear (RAL) estimator T_N can be constructed satisfying $T_N - \psi = \frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{V}_{(i)}; \psi, \eta) + o_P(N^{-1/2})$. When the IF can be decomposed as $\phi(\mathbf{V}; \psi, \eta) = \mathcal{V}(\mathbf{V}; \eta) - \psi$ for some function $\mathcal{V}(\mathbf{V}; \eta)$, called the *uncentered influence function (UIF)*, the corresponding RAL estimator is $T_N = \frac{1}{N} \sum_{i=1}^{N} \mathcal{V}(\mathbf{V}_{(i)}, \widehat{\eta})$ where $\widehat{\eta}$ denotes nuisances estimated from sample $\mathcal{D}$ (Kennedy, 2020a). We will focus on deriving UIFs in this paper. Once we have a UIF the corresponding IF could be expressed as $\phi(\mathbf{V}; \psi, \eta) = \mathcal{V}(\mathbf{V}; \eta) - \mathbb{E}_P[\mathcal{V}(\mathbf{V}; \eta)]$.

Double/Debiased Machine Learning (DML). DML methods (Chernozhukov et al., 2018) are based on two ideas: (1) use a *Neyman orthogonal score*[1] to estimate the target ψ, and (2) use *cross-fitting*[2] to construct the estimator. DML estimators guarantee $\sqrt{N}$-consistency even when the estimates $\widehat{\eta}$ of (possibly high-dimensional) nuisance functions converge at a much slower $N^{-1/4}$ rate ('*debiasedness*'), allowing the use of a broad array of modern ML methods that do not meet certain smoothness/complexity restrictions (i.e., *Donsker* class). Neyman-orthogonal scores may coincide with IFs - a fact we exploit in this paper.

[1] A Neyman orthogonal score is a function ϕ satisfying $\mathbb{E}_P[\phi(\mathbf{V}; \psi, \eta^*)] = 0$ and $\frac{\partial}{\partial \eta} \mathbb{E}_P[\phi(\mathbf{V}; \psi, \eta)]|_{\eta = \eta^*} = 0$, where η^* denotes the true nuisance.

[2] The cross-fitting technique uses distinct sets of samples in model training and estimator's evaluation.

3. IFs for Canonical Expressions

Before deriving IFs for any identifiable causal effects in PAGs, in this section, we derive IFs for two typical functionals that often appear in the expressions of causal effects, called here *canonical expressions*.

3.1. Canonical expression 1

Definition 3 (Canonical expression 1 (CE-1)). Let $\mathbf{T} = \{\mathbf{B}_1 < \cdots < \mathbf{B}_n\}$ be a set of ordered sets[3]. Let $\mathbf{C} \subseteq \mathbf{T}$ and $\mathbf{A}$ be a subset of variables contained in $\mathbf{C}$. A quantity $\mathcal{Q}$ is said to be (in the form of) a *canonical expression 1 (CE-1)* if it is in the following form:

$$\mathcal{Q} = \sum_{\mathbf{a}} \prod_{\mathbf{B}_i \in \mathbf{C}} P(\mathbf{b}_i|\text{pre}_{\mathbf{T}}(\mathbf{b}_i)). \tag{1}$$

For concreteness, we show the causal effect $P_{\mathbf{x}}(y)$ (for $\mathbf{X} = \{X_1, X_2\}$) in the PAG in Fig. 2a can be expressed as a CE-1 as follows:

Given a PTO $\mathbf{V} = \{C \prec B \prec A \prec X_1 \prec Z \prec X_2 \prec Y\}$, we have $Q[\mathbf{V} \setminus X_2]$ is identifiable from $Q[\mathbf{V}] = P(\mathbf{V})$ by Prop. 1 as X_2 is a bucket satisfying $\mathcal{C}(X_2) \cap Ch(X_2) = \{X_2\}$ and $\mathbf{S}_{X_2} = \{X_2\}$, and we obtain $Q[\mathbf{V} \setminus X_2] = P_{x_2}(\mathbf{v} \setminus x_2) = P(\mathbf{v})/P(x_2|\text{pre}(x_2)) = P(y|\text{pre}(y))P(\text{pre}(x_2))$. For $\mathbf{T} \equiv \mathbf{V}\setminus\{X_2\}$, $Q[\mathbf{T} \setminus X_1]$ is identifiable from $Q[\mathbf{T}]$ by Prop. 1 as X_1 is a bucket satisfying $\mathcal{C}(X_1) \cap Ch(X_1) = \{X_1\}$ and $\mathbf{S}_{X_1} = \{X_1\}$, and we obtain $Q[\mathbf{T} \setminus X_1] = P_{x_1,x_2}(\mathbf{t}\setminus\{x_1\}) = P_{x_2}(\mathbf{t})/P_{x_2}(x_1|\text{pre}_{\mathbf{T}}(x_1)) = P(y|\text{pre}(y))P(z|\text{pre}(z))P(a,b,c)$ by the equality $P_{x_2}(x_1|\text{pre}_{\mathbf{T}}(x_1)) = P(x_1|\text{pre}_{\mathbf{T}}(x_1))$. Finally, the causal effect $P_{\mathbf{x}}(y)$ is given as a CE-1 as:

$$P_{\mathbf{x}}(y) = \sum_{z,a,b,c} P(y|\text{pre}(y))P(z|\text{pre}(z))P(a|b,c)P(b|c)P(c). \tag{2}$$

We derive an IF for functionals in the form of CE-1 as follows:

Lemma 1 (UIF for CE-1). *Let a target estimand $\psi = \mathcal{Q}$ be a CE-1 given by Eq. (1) in Def. 3. Let $\mathbf{Y} \equiv \mathbf{C}\setminus\mathbf{A}$, and $\mathbf{X} \equiv \mathbf{T} \setminus \mathbf{C} \equiv \{\mathbf{B}_{j_1} < \cdots < \mathbf{B}_{j_m}\}$ where $\mathbf{B}_{j_s} \in \mathbf{T}$. Let $\mathbf{C}$ be partitioned with respect to $\mathbf{X}$ as $\mathbf{C} = \bigcup_{k=0}^{m} \mathbf{C}_k$, where $\mathbf{C}_k \equiv \{\mathbf{B}_r \in \mathbf{C} : j_k < r < j_{k+1}\} \equiv \{\mathbf{B}_{k_{\min}} < \cdots < \mathbf{B}_{k_{\max}}\}$ with $j_0 \equiv 0$ and $j_{m+1} \equiv n+1$. Let P_π be a distribution over $\mathbf{T}$ given by $P_\pi \equiv I_{\mathbf{x}}(\mathbf{X}) \prod_{\mathbf{B}_i \in \mathbf{C}} P(\mathbf{B}_i|pre_{\mathbf{T}}(\mathbf{B}_i))$. Then, $\mathcal{V}(\mathbf{T}; \eta = (\boldsymbol{\omega}, \boldsymbol{\theta}))$ in the following is a UIF for ψ:*

$$\mathcal{V}(\mathbf{T}; \eta = (\boldsymbol{\omega}, \boldsymbol{\theta})) = \theta_{0,1} + \sum_{\substack{k=1 \\ \mathbf{C}_k \neq \emptyset}}^{m} \omega_k \left(\theta_{k,1} - \theta_{k,2}\right), \tag{3}$$

where $\boldsymbol{\omega} \equiv \{\omega_k |\ \mathbf{C}_k \neq \emptyset,\ k \in \{1, \cdots, m\}\}$ and $\boldsymbol{\theta} \equiv \{\theta_{0,1}\} \cup \{(\theta_{k,1}, \theta_{k,2})|\ \mathbf{C}_k \neq \emptyset,\ k \in \{1, \cdots, m\}\}$ are nuisances given by $\omega_k \equiv \prod_{r=1}^{k} \frac{I_{\mathbf{b}_{j_r}}(\mathbf{B}_{j_r})}{P(\mathbf{B}_{j_r}|pre_{\mathbf{T}}(\mathbf{B}_{j_r}))}$, $\theta_{k,1} \equiv \mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|\mathbf{B}_{k_{\max}}, pre_{\mathbf{T}}(\mathbf{B}_{k_{\max}})\right]$, $\theta_{k,2} \equiv \mathbb{E}_{P_\pi}\left[I_{\mathbf{y}}(\mathbf{Y})|pre_{\mathbf{T}}(\mathbf{B}_{k_{\min}})\right]$ where $\theta_{0,1} = \mathbb{E}_{P_\pi}[I_{\mathbf{y}}(\mathbf{Y})]$ if $\mathbf{C}_0 = \emptyset$.

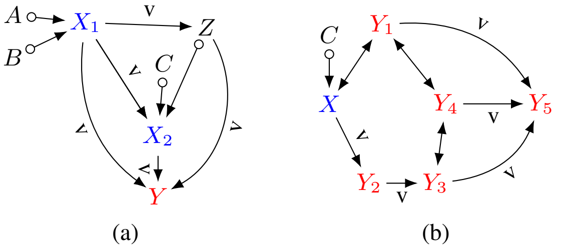

Figure 2: Example PAGs. Causal effects $P_{\mathbf{x}}(\mathbf{y})$ are identifiable and given by **(a)** CE-1, **(b)** CE-2.

For concreteness, we apply Lemma 1 to derive a UIF for $\psi \equiv P_{x_1,x_2}(y)$ in Fig. 2a which is identified as a CE-1 given in Eq. (2).

Illustration 1 (UIF for $P_{x_1,x_2}(y)$ in Fig. 2a). *Let $\mathbf{T} = \{C \prec B \prec A \prec X_1 \prec Z \prec X_2 \prec Y\}$, $\mathbf{C} = \{C \prec B \prec A \prec Z \prec Y\}$, and $\mathbf{X} = \{X_1 \prec X_2\}$. We have $\mathbf{C}_0 = \{C \prec B \prec A\}$, $\mathbf{C}_1 = \{Z\}$, and $\mathbf{C}_2 = \{Y\}$. Then Lemma 1 gives a UIF for ψ as*

$$\mathcal{V}_{P_{\mathbf{x}}(y)} = \theta_{0,1} + \omega_1\left(\theta_{1,1} - \theta_{1,2}\right) + \omega_2\left(\theta_{2,1} - \theta_{2,2}\right), \tag{4}$$

where $\omega_1 = \frac{I_{x_1}(X_1)}{P(X_1|pre(X_1))}$, $\omega_2 = \frac{I_{x_1,x_2}(X_1,X_2)}{P(X_1|pre(X_1))P(X_2|pre(X_2))}$; for $P_\pi \equiv I_{x_1,x_2}(X_1,X_2)P(A,B,C)P(Z|pre(Z))P(Y|pre(Y))$, $\theta_{0,1} = \mathbb{E}_{P_\pi}\left[I_y(Y)|pre(X_1)\right]$, $\theta_{1,1} = \mathbb{E}_{P_\pi}\left[I_y(Y)|pre(X_2)\right]$, $\theta_{1,2} = \mathbb{E}_{P_\pi}\left[I_y(Y)|pre(Z)\right]$; and $\theta_{2,1} = \mathbb{E}_{P_\pi}\left[I_y(Y)|\mathbf{T}\right] = I_y(Y)$ and $\theta_{2,2} = \mathbb{E}_{P_\pi}\left[I_y(Y)|pre(Y)\right]$.

3.2. Canonical expression 2

Definition 4 (Canonical expression 2 (CE-2)). Let $\mathcal{Q}_1$ and $\mathcal{Q}_2$ be two CE-1s, then the quantity $\mathcal{Q} = \sum_{\mathbf{z}}(\mathcal{Q}_1 \times \mathcal{Q}_2)$ is said to be (in the form of) a *canonical expression 2 (CE-2)*.

A broad class of causal effects are identified as a CE-2, including all joint interventional distributions ($P_x(\mathbf{v})$) when X is singleton (Jaber et al., 2018b, Thm. 1), as well as in the following scenario which follows from Prop. 1:

Corollary 1. *Let a PTO in PAG $\mathcal{G}$ over $\mathbf{V}$ be $\mathbf{B}_1 \prec \cdots \prec \mathbf{B}_m$. Let $\mathbf{X}, \mathbf{Y} \subset \mathbf{V}$ with $\mathbf{X}$ being a bucket. Then, if $\mathcal{C}(\mathbf{X}) \cap Ch(\mathbf{X}) \subseteq \mathbf{X}$, $P_{\mathbf{x}}(\mathbf{y})$ is identifiable and given by*

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{v}\setminus(\mathbf{x}\cup\mathbf{y})} \mathcal{Q}_{\mathbf{V}\setminus\mathbf{S}_{\mathbf{X}}} \times \mathcal{Q}_{\mathbf{S}_{\mathbf{X}}\setminus\mathbf{X}}, \tag{5}$$

[3] We use $\mathbf{W} = \{\mathbf{B}_1 < \cdots < \mathbf{B}_k\}$ to denote a set of ordered sets $\mathbf{W} = \{\mathbf{B}_1, \cdots, \mathbf{B}_k\}$ or a union of ordered sets $\mathbf{W} = \cup_{i=1}^{k}\mathbf{B}_i$ depending on the context.

where $\mathcal{Q}_{\mathbf{V}\setminus\mathbf{S_X}} \equiv \prod_{\mathbf{B}_i\subseteq\mathbf{V}\setminus\mathbf{S_X}} P(\mathbf{b}_i|\textit{pre}(\mathbf{b}_i))$*,* $\mathcal{Q}_{\mathbf{S_X}\setminus\mathbf{X}} \equiv \sum_{\mathbf{x}}\prod_{\mathbf{B}_i\subseteq\mathbf{S_X}} P(\mathbf{b}_i|\textit{pre}(\mathbf{b}_i))$*, and* $\mathbf{S_X} = \bigcup_{X\in\mathbf{X}}\mathbf{S}_X$ *with* $\mathbf{S}_X$ *being the DC-component of* X*.*

Eq. (5) is a CE-2 where $\mathcal{Q}_{\mathbf{V}\setminus\mathbf{S_X}}$ and $\mathcal{Q}_{\mathbf{S_X}\setminus\mathbf{X}}$ are CE-1s. As a concrete example, consider the PAG in Fig. 2b with a PTO $\mathbf{V} = \{C \prec X \prec Y_1 \prec Y_2 \prec Y_3 \prec Y_4 \prec Y_5\}$. Since X is a bucket and satisfies $\mathcal{C}(X) \cap Ch(X) = \{X\}$ with $\mathcal{C}(X) = \{X, C, Y_1, Y_4, Y_3\}$ and $Ch(X) = \{X, Y_2\}$, the causal effect $P_x(\mathbf{y})$ where $\mathbf{Y} = \{Y_1, \cdots, Y_5\}$ is identifiable by Coro. 1 and given by

$$P_x(\mathbf{y}) = \sum_c \mathcal{Q}_{\mathbf{V}\setminus\mathbf{S}_X}\mathcal{Q}_{\mathbf{S}_X\setminus X}, \tag{6}$$

where $\mathbf{S}_X = \{X, Y_1, Y_3, Y_4\}$, $\mathbf{V}\setminus\mathbf{S}_X = \{C, Y_2, Y_5\}$, $\mathcal{Q}_{\mathbf{V}\setminus\mathbf{S}_X} \equiv P(y_5|\text{pre}(y_5))P(y_2|\text{pre}(y_2))P(c)$, and $\mathcal{Q}_{\mathbf{S}_X\setminus X} \equiv \sum_{x'} P(y_3, y_4|y_1, y_2, x', c)P(y_1, x'|c)$.

We derive an IF for CE-2 as follows:

Lemma 2 **(UIF for CE-2).** *Let a target estimand* $\psi = \mathcal{Q}$ *be a CE-2 given in Def. 4. Let* $\mathcal{V}_i$ *be a UIF for the CE-1* $\mathcal{Q}_i$ *given in Lemma 1 and* $\mu_i \equiv \mathbb{E}_P[\mathcal{V}_i]$ *for* $i \in \{1, 2\}$*. Then,* $\mathcal{V}(\mathbf{V}; \eta)$ *below is a UIF for* ψ*:*

$$\mathcal{V}(\mathbf{V}; \eta) = \sum_{\mathbf{z}}(\mathcal{V}_1\mu_2 + (\mathcal{V}_2 - \mu_2)\mu_1). \tag{7}$$

Lemma 2 provides a UIF for any causal effects that are identifiable by Coro. 1. For a concrete example, we will use Lemma 2 to derive a UIF for $\psi \equiv P_x(\mathbf{y})$ in Fig. 2b identified by Coro. 1 as given in Eq. (6).

Illustration 2 **(UIF for $P_x(\mathbf{y})$ in Fig. 2b).** *A UIF for* $P_x(\mathbf{y})$ *in Eq. (6) is given by Lemma 2 as*

$$\mathcal{V}_{P_x(\mathbf{y})} = \sum_c \left(\mathcal{V}_{\mathbf{V}\setminus\mathbf{S_X}}\mu_{\mathbf{S_X}\setminus\mathbf{X}} + (\mathcal{V}_{\mathbf{S_X}\setminus\mathbf{X}} - \mu_{\mathbf{S_X}\setminus\mathbf{X}})\mu_{\mathbf{V}\setminus\mathbf{S_X}}\right), \tag{8}$$

where $\mathcal{V}_{\mathbf{V}\setminus\mathbf{S_X}}$ *is a UIF for* $\mathcal{Q}_{\mathbf{V}\setminus\mathbf{S}_X}$ *and, by Lemma 1, is given with* $\mathbf{V} = \{C \prec X \prec Y_1 \prec Y_2 \prec Y_3 \prec Y_4 \prec Y_5\}$ *as* $\mathcal{V}_{\mathbf{V}\setminus\mathbf{S_X}} = \theta^a_{0,1} + \omega^a_1(\theta^a_{1,1} - \theta^a_{1,2}) + \omega^a_2(\theta^a_{2,1} - \theta^a_{2,2})$*, where* $\omega^a_1 = \frac{I_{x,y_1}(X,Y_1)}{P(X|C)P(Y_1|X,C)}$ *and* $\omega^a_2 = \omega^a_1 \times \frac{I_{y_3,y_4}(Y_3,Y_4)}{P(Y_3|\textit{pre}(Y_3))P(Y_4|\textit{pre}(Y_4))}$*; and for* $P_{\pi^a} \equiv I_{x,y_1,y_3,y_4}(X, Y_1, Y_3, Y_4)P(C)P(Y_2|\textit{pre}(Y_2))P(Y_5|\textit{pre}(Y_5))$ *and* $I^a \equiv I_{c,y_2,y_5}(C, Y_2, Y_5)$*,* $\theta^a_{0,1} = \mathbb{E}_{P_{\pi^a}}[I^a|C]$*,* $\theta^a_{1,1} = \mathbb{E}_{P_{\pi^a}}[I^a|Y_2, \textit{pre}(Y_2)]$*,* $\theta^a_{1,2} = \mathbb{E}_{P_{\pi^a}}[I^a|\textit{pre}(Y_2)]$*,* $\theta^a_{2,1} = I^a$*, and* $\theta^a_{2,2} = \mathbb{E}_{P_{\pi^a}}[I^a|\textit{pre}(Y_5)]$*.*

Also, $\mathcal{V}_{\mathbf{S_X}\setminus\mathbf{X}}$ *is a UIF for* $\mathcal{Q}_{\mathbf{S_X}\setminus\mathbf{X}}$ *and is given by Lemma 1 as* $\mathcal{V}_{\mathbf{S_X}\setminus\mathbf{X}} = \theta^b_{0,1} + \omega^b_1(\theta^b_{1,1} - \theta^b_{1,2}) + \omega^b_2(\theta^b_{2,1} - \theta^b_{2,2})$*, where* $\omega^b_1 = \frac{I_c(C)}{P(C)}$ *and* $\omega^b_2 = \omega^b_1 \times \frac{I_{y_2}(Y_2)}{P(Y_2|\textit{pre}(Y_2))}$*; and for* $P_{\pi^b} \equiv I_{c,y_2}(C, Y_2)P(Y_3, Y_4|\textit{pre}(Y_3))P(X, Y_1|C)$ *and* $I^b \equiv I_{y_1,y_3,y_4}(Y_1, Y_3, Y_4)$*,* $\theta^b_{0,1} = \mathbb{E}_{P_{\pi^b}}[I^b]$*,* $\theta^b_{1,1} = \mathbb{E}_{P_{\pi^b}}[I^b|Y_1, \textit{pre}(Y_1)]$*,* $\theta^b_{1,2} = \mathbb{E}_{P_{\pi^b}}[I^b|\textit{pre}(X)]$*,* $\theta^b_{2,1} = I^b$*, and* $\theta^b_{2,2} = \mathbb{E}_{P_{\pi^2}}[I^b|\textit{pre}(Y_3)]$*.*

Finally $\mu_{\mathbf{V}\setminus\mathbf{S_X}} \equiv \mathbb{E}_P[\mathcal{V}_{\mathbf{V}\setminus\mathbf{S_X}}]$*, and* $\mu_{\mathbf{S_X}\setminus\mathbf{X}} \equiv \mathbb{E}_P[\mathcal{V}_{\mathbf{S_X}\setminus\mathbf{X}}]$*. Refer Appendix A for derivation details.*

4. IFs for Causal Estimands

In this section, we derive IFs for any identifiable causal effects in PAGs, armed with IFs for the canonical expressions discussed in the previous section. We develop a complete algorithm for deriving IFs by recursively deriving IFs of C-factors $Q[\cdot]$ inspired by IDP algorithm (Jaber et al., 2019) which recursively identifies C-factors by repeated application of Prop. 1 or 2. We will first develop basic results for deriving IFs of C-factors corresponding to Prop. 1 and 2.

Prop. 1 computes $Q\left[\mathbf{T}\setminus\mathbf{X}\right]$ in terms of given $Q\left[\mathbf{T}\right]$. We first rewrite Prop. 1 in a form more amenable for the purpose of deriving IFs:

Lemma 3. *Let* G *be a PAG over* $\mathbf{V}$*,* $\mathbf{T} = \cup_{i=1}^m \mathbf{B}_i$ *be the union of a set of buckets, and* $\mathbf{X} \subseteq \mathbf{T}$ *be a bucket. Given* $Q[\mathbf{T}]$ *and a PTO* $\mathbf{B}_1 \prec \cdots \prec \mathbf{B}_m$ *with respect to* $G(\mathbf{T})$*,* $Q\left[\mathbf{T}\setminus\mathbf{X}\right]$ *is identifiable if and only if* $\mathcal{C}(\mathbf{X}) \cap Ch(\mathbf{X}) \subseteq \mathbf{X}$ *in* $G(\mathbf{T})$*. When* $Q\left[\mathbf{T}\setminus\mathbf{X}\right]$ *is identifiable, letting* $\mathbf{S_X} = \bigcup_{X\in\mathbf{X}}\mathbf{S}_X$ *with* $\mathbf{S}_X$ *being the DC-component of* X *in* $G(\mathbf{T})$*, then* $\mathbf{S_X}$ *consists of a union of buckets. Denoting* $\mathbf{S_X} = \{\mathbf{B}_{j_1}, \cdots, \mathbf{B}_{j_p}\}$ *and* $\mathbf{T}\setminus\mathbf{S_X} = \{\mathbf{B}_{i_1}, \cdots, \mathbf{B}_{i_q}\}$*,* $Q\left[\mathbf{T}\setminus\mathbf{X}\right]$ *is given by*

$$Q\left[\mathbf{T}\setminus\mathbf{X}\right] = \mathcal{Q}_{\mathbf{T}\setminus\mathbf{S_X}} \times \mathcal{Q}_{\mathbf{S_X}\setminus\mathbf{X}}, \tag{9}$$

where $\mathcal{Q}_{\mathbf{T}\setminus\mathbf{S_X}} \equiv \prod_{\mathbf{B}_{i_r}\in\mathbf{T}\setminus\mathbf{S_X}} P_{\mathbf{v}\setminus\mathbf{t}}(\mathbf{b}_{i_r}|\textit{pre}_{\mathbf{T}}(\mathbf{b}_{i_r}))$*, and* $\mathcal{Q}_{\mathbf{S_X}\setminus\mathbf{X}} \equiv \sum_{\mathbf{x}}\prod_{\mathbf{B}_{j_s}\in\mathbf{S_X}} P_{\mathbf{v}\setminus\mathbf{t}}(\mathbf{b}_{j_s}|\textit{pre}_{\mathbf{T}}(\mathbf{b}_{j_s}))$*.*

For any $\mathbf{W} \subseteq \mathbf{V}$, we will use $\phi_{Q[\mathbf{W}]}$ to denote an IF for the C-factor $Q[\mathbf{W}]$, $\mathcal{V}_{Q[\mathbf{W}]}$ the corresponding UIF, and $\mu_{Q[\mathbf{W}]} \equiv \mathbb{E}_P[\mathcal{V}_{Q[\mathbf{W}]}]$. We derive an IF for $Q\left[\mathbf{T}\setminus\mathbf{X}\right]$ that is identified by Lemma 3 in terms of $\mathcal{V}_{Q[\mathbf{T}]}$ as follows:

Lemma 4 **(IF of C-factors).** *Suppose* $\psi \equiv Q\left[\mathbf{T}\setminus\mathbf{X}\right]$ *is identifiable via Lemma 3 and given by Eq. (9). Then, given* $\mathcal{V}_{Q[\mathbf{T}]}$*,* $\mathcal{V} \equiv \mathcal{V}_{Q[\mathbf{T}\setminus\mathbf{X}]}$ *below is a UIF for* ψ*:*

$$\mathcal{V} = \mathcal{V}_{\mathbf{S_X}\setminus\mathbf{X}}\mu_{\mathcal{V}_{\mathbf{T}\setminus\mathbf{S_X}}} + (\mathcal{V}_{\mathbf{T}\setminus\mathbf{S_X}} - \mu_{\mathcal{V}_{\mathbf{T}\setminus\mathbf{S_X}}})\mu_{\mathbf{S_X}\setminus\mathbf{X}}, \tag{10}$$

where $(\mathcal{V}_{\mathbf{S_X}\setminus\mathbf{X}}, \mathcal{V}_{\mathbf{T}\setminus\mathbf{S_X}})$ *are UIFs for* $(\mathcal{Q}_{\mathbf{S_X}\setminus\mathbf{X}}, \mathcal{Q}_{\mathbf{T}\setminus\mathbf{S_X}})$ *respectively, given by*

$$\mathcal{V}_{\mathbf{S_X}\setminus\mathbf{X}} \equiv \sum_{\mathbf{x}}(\mathcal{V}_{j_1}\prod_{k=2}^{p}\mu_{j_k} + \sum_{k=2}^{p}\phi_{j_k}\prod_{\ell=1,\ell\neq k}^{p}\mu_{j_\ell}),$$

$$\mathcal{V}_{\mathbf{T}\setminus\mathbf{S_X}} \equiv \mathcal{V}_{i_1}\prod_{r=2}^{q}\mu_{i_r} + \sum_{r=2}^{q}\phi_{i_r}\prod_{\ell=1,\ell\neq r}^{q}\mu_{i_\ell},$$

where, for $c \in \{1, 2, \cdots, m\}$*,* $\mathcal{V}_c \equiv \frac{\sum_{\mathbf{t}\setminus\{\mathbf{b}_c,\textit{pre}_{\mathbf{T}}(\mathbf{b}_c)\}}\mathcal{V}_{Q[\mathbf{T}]}}{\sum_{\mathbf{t}\setminus\textit{pre}_{\mathbf{T}}(\mathbf{b}_c)}\mu_{Q[\mathbf{T}]}} -$

$\frac{\sum_{\mathbf{t}\setminus\{\mathbf{b}_c,pre_{\mathbf{T}}(\mathbf{b}_c)\}}\mu_{Q[\mathbf{T}]}}{\sum_{\mathbf{t}\setminus pre_{\mathbf{T}}(\mathbf{b}_c)}\mu_{Q[\mathbf{T}]}} \cdot \frac{\sum_{\mathbf{t}\setminus pre_{\mathbf{T}}(\mathbf{b}_c)}\phi_{Q[\mathbf{T}]}}{\sum_{\mathbf{t}\setminus pre_{\mathbf{T}}(\mathbf{b}_c)}\mu_{Q[\mathbf{T}]}}$, $\mu_c \equiv \mathbb{E}_P[\mathcal{V}_c]$, *and* $\phi_c \equiv \mathcal{V}_c - \mu_c$.

The following lemma derives an IF for the C-factor $Q[\mathbf{C}]$ from the IFs of C-factors over some subsets of C, corresponding to the C-factor decomposition in Prop. 2.

Lemma 5 (Decomposition of IFs). *For* $\mathbf{A} \subseteq \mathbf{C} \subseteq \mathbf{V}$,

$$\mathcal{V}_{Q[\mathbf{C}]} = (a) + (b) - (c), \tag{11}$$

where $(a) = \frac{\mathcal{V}_{Q[\mathcal{R}_{\mathbf{A}}]}\cdot\mu_{Q[\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{A}}}]}}{\mu_{Q[\mathcal{R}_{\mathbf{A}}\cap\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{A}}}]}}$, $(b) = \frac{\mu_{Q[\mathcal{R}_{\mathbf{A}}]}\cdot\phi_{Q[\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{A}}}]}}{\mu_{Q[\mathcal{R}_{\mathbf{A}}\cap\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{A}}}]}}$, $(c) = \frac{\mu_{Q[\mathcal{R}_{\mathbf{A}}]}\cdot\mu_{Q[\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{A}}}]}}{\mu_{Q[\mathcal{R}_{\mathbf{A}}\cap\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{A}}}]}} \cdot \frac{\phi_{Q[\mathcal{R}_{\mathbf{A}}\cap\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{A}}}]}}{\mu_{Q[\mathcal{R}_{\mathbf{A}}\cap\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{A}}}]}}$ *with* $\mathcal{R}_{(\cdot)} = \mathcal{R}^{\mathbf{C}}_{(\cdot)}$.

Finally, we develop a systematic procedure named IFP, given in Algo. 1, that derives a UIF for any identifiable causal effect in PAGs. IFP recursively applies Lemmas 4 and 5 until all needed C-factors are in CE-1 or CE-2 form, whose UIFs are given by Lemma 1 and 2, respectively, initially equipped with a UIF for $P(\mathbf{v})$, $\mathcal{V}_{Q[\mathbf{V}]} = I_{\mathbf{v}}(\mathbf{V})$.

Theorem 1 (Completeness of IFP). *Procedure IFP (Algo. 1) derives a UIF for any identifiable* $P_{\mathbf{x}}(\mathbf{y})$ *in a PAG* G *over* $\mathbf{V}$ *in* $O(|\mathbf{V}|^4)$ *time, where* $|\mathbf{V}|$ *is the number of variables. IFP returns* FAIL *if* $P_{\mathbf{x}}(\mathbf{y})$ *is not identifiable.*

Algorithm 1 IFP$(\mathbf{x}, \mathbf{y}, G(\mathbf{V}), P)$

```
1: Input: Two disjoint sets $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$; A PAG $G$ over $\mathbf{V}$; A distribution $P(\mathbf{v})$.
2: Output: Expression for UIF $\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}$ or FAIL.
3: Let $\mathbf{D} = An(\mathbf{Y})_{G(\mathbf{V}\setminus\mathbf{X})}$.
4: $\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})} = \sum_{\mathbf{d}\setminus\mathbf{y}} \text{DeriveUIF}\left(\mathbf{D}, \mathbf{V}, P(\mathbf{V}), \mathcal{V}_{Q[\mathbf{V}]} = I_{\mathbf{v}}(\mathbf{V})\right)$
5: function DeriveUIF $(\mathbf{C}, \mathbf{T}, Q = Q[\mathbf{T}], \mathcal{V} = \mathcal{V}_Q)$
6:     if $\mathbf{C} = \emptyset$, then return $1$.
7:     if $\mathbf{C} = \mathbf{T}$, then return $\mathcal{V}$.
       {$\mathbf{B}$ denotes a bucket in $G(\mathbf{T})$; $\mathcal{C}(\mathbf{B})$ the PC-component of $\mathbf{B}$ in $G(\mathbf{T})$, and $\mathcal{R}_{(\cdot)} \equiv \mathcal{R}^{\mathbf{C}}_{(\cdot)}$.}
8:     if $\exists\mathbf{B} \subseteq \mathbf{T}\setminus\mathbf{C}$ s.t. $\mathcal{C}(\mathbf{B}) \cap Ch(\mathbf{B}) \subseteq \mathbf{B}$, then
9:         Compute $Q[\mathbf{T}\setminus\mathbf{B}]$ from $Q$ via Lemma 3.
10:        if $Q[\mathbf{T}\setminus\mathbf{B}]$ is expressible as CE-1,
             then, Compute $\mathcal{V}_{Q[\mathbf{T}\setminus\mathbf{B}]}$ via Lemma 1.
11:        else if $Q[\mathbf{T}\setminus\mathbf{B}]$ is expressible as CE-2,
             then, Compute $\mathcal{V}_{Q[\mathbf{T}\setminus\mathbf{B}]}$ via Lemma 2.
12:        else, Compute $\mathcal{V}_{Q[\mathbf{T}\setminus\mathbf{B}]}$ via Lemma 4.
13:        return DeriveUIF $\left(\mathbf{C}, \mathbf{T}\setminus\mathbf{B}, Q[\mathbf{T}\setminus\mathbf{B}], \mathcal{V}_{Q[\mathbf{T}\setminus\mathbf{B}]}\right)$.
14:    else if $\exists\mathbf{B} \subseteq \mathbf{C}$ s.t. $\mathcal{R}_{\mathbf{B}} \neq \mathbf{C}$, then
15:        return $(a) + (b) - (c)$, where
       { Let UIF$(\mathbf{W})$ = DeriveUIF$(\mathbf{W}, \mathbf{T}, Q, \mathcal{V})$; IF$(\mathbf{W})$ = UIF$(\mathbf{W}) - \mathbb{E}_P[$UIF$(\mathbf{W})]$; ID$(\mathbf{W}) = \mathbb{E}_P[$UIF$(\mathbf{W})]$}
       $(a) = \frac{\text{UIF}(\mathcal{R}_{\mathbf{B}})\cdot\text{ID}(\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{B}}})}{\text{ID}(\mathcal{R}_{\mathbf{B}}\cap\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{B}}})}$; $(b) = \frac{\text{IF}(\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{B}}})\cdot\text{ID}(\mathcal{R}_{\mathbf{B}})}{\text{ID}(\mathcal{R}_{\mathbf{B}}\cap\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{B}}})}$;
       $(c) = \frac{\text{ID}(\mathcal{R}_{\mathbf{B}})\cdot\text{ID}(\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{B}}})}{\text{ID}(\mathcal{R}_{\mathbf{B}}\cap\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{B}}})} \cdot \frac{\text{IF}(\mathcal{R}_{\mathbf{B}}\cap\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{B}}})}{\text{ID}(\mathcal{R}_{\mathbf{B}}\cap\mathcal{R}_{\mathbf{C}\setminus\mathcal{R}_{\mathbf{B}}})}$.
16:    else return FAIL.
17: end function
```

For concreteness, we demonstrate the application of IFP by deriving a UIF for $\psi = P_x(\mathbf{y})$, where $\mathbf{Y} \equiv \{Y_1, Y_2, Y_3, Y_4\}$, in the PAG in Fig. 1.

Illustration 3 (UIF for $P_x(\mathbf{y})$ in Fig. 1 by IFP). *We start with* $\mathbf{D} \equiv \mathbf{Y}$ *(Line 3) and* $\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})} =$ DeriveUIF$(\mathbf{D}, \mathbf{V}, P(\mathbf{V}), \mathcal{V}_{Q[\mathbf{V}]} = I_{\mathbf{v}}(\mathbf{V}))$ *(Line 4).* DeriveUIF() *reaches line 14, where* $\mathbf{B}_0 \equiv \{Y_2\}$ *satisfies the condition with* $\mathcal{R}_{\mathbf{B}_0} = \{Y_2, Y_3\}$, $\mathcal{R}_{\mathbf{D}\setminus\mathcal{R}_{\mathbf{B}_0}} = \{Y_1, Y_4\}$, *and* $\mathcal{R}_{\mathbf{B}_0} \cap \mathcal{R}_{\mathbf{D}\setminus\mathcal{R}_{\mathbf{B}_0}} = \emptyset$. *Then, line 15 gives (using* ID$(\emptyset) = 1$ *and* IF$(\emptyset) = 0$*)*

$$\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})} = \text{UIF}(\mathcal{R}_{\mathbf{B}_0}) \cdot \text{ID}(\mathcal{R}_{\mathbf{D}\setminus\mathcal{R}_{\mathbf{B}_0}}) + \text{IF}(\mathcal{R}_{\mathbf{D}\setminus\mathcal{R}_{\mathbf{B}_0}}) \cdot \text{ID}(\mathcal{R}_{\mathbf{B}_0}).$$

Next we show a sketch derivation of UIF$(\mathcal{R}_{\mathbf{B}_0})$ = DeriveUIF$(\mathcal{R}_{\mathbf{B}_0}, \mathbf{V}, P(\mathbf{V}), I_{\mathbf{v}}(\mathbf{V}))$. *(Refer Appendix A for details).* UIF$(\mathcal{R}_{\mathbf{B}_0})$ *is derived by repeating Lines 8, 9, 10, and 13 as follows: Starting with* $\mathbf{B} = Y_4$ *at Line 8, let* $\mathbf{T} = \mathbf{V} \setminus \mathbf{B} = \{Y_1, R, X, Y_2, Y_3\}$, *compute* $Q[\mathbf{T}]$ *(Line 9) and* $\mathcal{V}_{Q[\mathbf{T}]}$ *(Line 10), call* DeriveUIF$(\mathcal{R}_{\mathbf{B}_0}, \mathbf{T}, Q[\mathbf{T}], \mathcal{V}_{Q[\mathbf{T}]})$ *(Line 13). Then repeat the above by calling* DeriveUIF$(\mathcal{R}_{\mathbf{B}_0}, \mathbf{T}, Q[\mathbf{T}], \mathcal{V}_{Q[\mathbf{T}]})$ *three more times with* $\mathbf{B} = Y_1$ *at line 8,* $\mathbf{T} = \{R, X, Y_2, Y_3\}$; $\mathbf{B} = X$ *at line 8,* $\mathbf{T} = \{R, Y_2, Y_3\}$; *and* $\mathbf{B} = R$ *at line 8,* $\mathbf{T} = \{Y_2, Y_3\}$. *Finally we obtain* $Q[\mathcal{R}_{\mathbf{B}_0}] = Q[Y_2, Y_3] = \sum_r P(y_2, y_3|x, r)P(r)$, *and* UIF$(\mathcal{R}_{\mathbf{B}_0}) = \mathcal{V}_{Q[\mathcal{R}_{\mathbf{B}_0}]}$ *is given by Lemma 1 as* UIF$(\mathcal{R}_{\mathbf{B}_0}) = \theta^a_{0,1} + \omega^a_1(\theta^a_{1,1} - \theta^a_{2,1})$, *where* $\omega^a_1 = \frac{I_x(X)}{P(X|R)}$; *and for* $P_{\pi^a} = I_x(X)P(Y_2, Y_3|X, R)P(R)$, $\theta^a_{0,1} = \mathbb{E}_{P_{\pi^a}}[I^a|R]$, $\theta^a_{1,1} = I^a$, *and* $\theta^a_{1,2} = \mathbb{E}_{P_{\pi^a}}[I^a|X, R]$ *where* $I^a \equiv I_{y_2,y_3}(Y_2, Y_3)$.

UIF$(\mathcal{R}_{\mathbf{D}\setminus\mathcal{R}_{\mathbf{B}_0}})$ = DeriveUIF$(\mathcal{R}_{\mathbf{D}\setminus\mathcal{R}_{\mathbf{B}_0}}, \mathbf{V}, P(\mathbf{V}), I_{\mathbf{v}}(\mathbf{V}))$ *is derived in a similar manner. We obtain* $Q[\mathcal{R}_{\mathbf{D}\setminus\mathcal{R}_{\mathbf{B}_0}}] = Q[Y_1, Y_4] = P(y_4|pre(y_4))P(y_1)$ *for PTO* $Y_1 \prec R \prec X \prec Y_2 \prec Y_3 \prec Y_4$, *and* UIF$(\mathcal{R}_{\mathbf{D}\setminus\mathbf{B}_0}) = \theta^b_{0,1} + \omega^b_1(\theta^b_{1,1} - \theta^b_{2,1})$ *where* $\omega^b_1 = \frac{I_{r,x,y_2,y_3}(R,X,Y_2,Y_3)}{P(R,X,Y_2,Y_3|Y_1)}$; *and for* $P_{\pi^b} = I_{r,x,y_2,y_3}(R, X, Y_2, Y_3)P(Y_4|pre(Y_4))P(Y_1)$, $\theta^b_{0,1} = \mathbb{E}_{P_{\pi^b}}[I^b|Y_1]$, $\theta^a_{1,1} = I^b$, *and* $\theta^b_{1,2} = \mathbb{E}_{P_{\pi^b}}[I^b|pre(Y_4)]$ *where* $I^b \equiv I_{y_1,y_4}(Y_1, Y_4)$.

For reference, $P_x(\mathbf{y})$ *is identified as*

$$P_x(\mathbf{y}) = Q[\mathbf{Y}] = Q[Y_2, Y_3]Q[Y_1, Y_4], \tag{12}$$

where $Q[Y_2, Y_3] = \sum_r P(y_2, y_3|x, r)P(r)$ *and* $Q[Y_1, Y_4] = P(y_4|pre(y_4))P(y_1)$.

5. DML Estimators

In this section, we construct DML estimators for causal effects $P_{\mathbf{x}}(\mathbf{y})$ from finite samples $\mathcal{D} = \{\mathbf{V}_{(i)}\}_{i=1}^N$ based on the UIF $\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V}; \eta)$ derived by IFP algorithm. The resulting DML estimators have nice properties of debiasedness, as well as doubly robustness in the sense that an estimator T_N composed of the nuisances $\eta = (\eta_0, \eta_1)$ is said to be *doubly robust* if T_N is consistent whenever either η_0 or η_1 are consistent.

First we show that IFs derived by IFP are a Neyman orthogonal score, which is needed for the DML method.

Proposition 3. *Let $P_{\mathbf{x}}(\mathbf{y})$ be identified as $P_{\mathbf{x}}(\mathbf{y}) = \psi \equiv \Psi(P)$. Then, the IF $\phi_{P_{\mathbf{x}}(\mathbf{y})} = \mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})} - \mathbb{E}_P[\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}]$, where $\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}$ is derived by Algo. 1 IFP, is a Neyman orthogonal score for ψ.*

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

Definition 5 (Double/Debiased Machine Learning estimator for identifiable causal effects (DML-IDP)). Let $\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V};\eta)$ be the UIF given by Algo. 1 IFP for the target functional $\psi = P_{\mathbf{x}}(\mathbf{y})$. Let $\mathcal{D} = \{\mathbf{V}_{(i)}\}_{i=1}^{N}$ denote samples drawn from $P(\mathbf{v})$. Then, the DML-IDP estimator T_N for $\psi = P_{\mathbf{x}}(\mathbf{y})$ is constructed as follows:
(1) Split $\mathcal{D}$ randomly into two halves: $\mathcal{D}_0$ and $\mathcal{D}_1$;
(2) For $p \in \{0,1\}$, use $\mathcal{D}_p$ to construct models for $\widehat{\eta}_p$, the nuisance functions estimated from samples $\mathcal{D}_p$; and
(3) $T_N \equiv \sum_{p\in\{0,1\}} \frac{2}{N} \sum_{\mathbf{V}_{(i)}\in\mathcal{D}_p} \mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V}_{(i)}, \widehat{\eta}_{1-p})$.

To witness the robustness properties of DML-IDP, we first note that the nuisances in $\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V};\eta)$ returned by IFP consist of the nuisances of UIFs for CE-1:

Lemma 6 (Nuisances of UIFs). *The UIF $\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V};\eta)$ returned by Algo. 1 IFP is an arithmetic combination (ratio, multiplication, and marginalization) of UIFs for functionals in the form of CE-1, denoted as $\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V};\eta = \{\boldsymbol{\omega}_j, \boldsymbol{\theta}_j\}_{j=1}^{\ell}) = \mathcal{A}\left(\{\mathcal{V}_j(\boldsymbol{\omega}_j, \boldsymbol{\theta}_j)\}_{j=1}^{\ell}\right)$ where $\mathcal{V}_j(\boldsymbol{\omega}_j, \boldsymbol{\theta}_j)$ denotes a UIF given by Lemma 1 with $\boldsymbol{\omega}_j = \{\omega_{j,k}\}_{k=1}^{m_j}$ and $\boldsymbol{\theta}_j = \{\theta_{j,0,1}\} \cup \{\theta_{j,k,1}, \theta_{j,k,2}\}_{k=1}^{m_j}$ being nuisances for $\mathcal{V}_j$, and $\mathcal{A}(\cdot)$ an arithmetic function.*

For example, the UIF for $P_x(\mathbf{y})$ in Fig. 2b given by Eq. (8) is a function of UIFs $\mathcal{V}_{\mathbf{V}\setminus\mathbf{S_x}}$ and $\mathcal{V}_{\mathbf{S_x}\setminus\mathbf{X}}$ both of which are given by Lemma 1 as shown in Illustration 2.

We show that DML-IDP estimators attain debiasedness and doubly robustness, the main result of this section:

Theorem 2 (Properties of DML-IDP). *Let T_N be the DML-IDP estimator of $P_{\mathbf{x}}(\mathbf{y})$ defined in Def. 5 constructed based on the UIF $\mathcal{V}_{P_{\mathbf{x}}(\mathbf{y})}(\mathbf{V};\eta = \{\boldsymbol{\omega}_j, \boldsymbol{\theta}_j\}_{j=1}^{\ell})$ where $\boldsymbol{\omega}_j = \{\omega_{j,k}\}_{k=1}^{m_j}$ and $\boldsymbol{\theta}_j = \{\theta_{j,0,1}\} \cup \{\theta_{j,k,1}, \theta_{j,k,2}\}_{k=1}^{m_j}$ are nuisances as specified in Lemma 6. Then,*

1. ***Debiasedness**: T_N is $\sqrt{N}$-consistent and asymptotically normal if estimates for all nuisances converge to the true nuisances at least at rate $o_P(N^{-1/4})$.*

2. ***Doubly robustness**: T_N is consistent if, for every $j = 1, \cdots, \ell$ and $k = 1, \cdots, m_j$, either estimates $\widehat{\omega}_{j,k}$ or $(\widehat{\theta}_{j,k-1,1}, \widehat{\theta}_{j,k,2})$ converge to the true nuisances at rate $o_P(1)$.*

By Thm. 2, DML-IDP estimators attain root-N consistency even when nuisances converge much slower at fourth-root-N rate or when some nuisances are misspecified. These properties allow one to employ flexible ML models (e.g., neural nets) that do not meet certain complexity restrictions (e.g., Donsker condition) for estimating nuisances in estimating causal effects. In contrast, plug-in estimators may fail to achieve $\sqrt{N}$-consistency if estimates for nuisances converges at $o_P(N^{-1/4})$ and are vulnerable to model misspecification.

For concreteness, we compare DML-IDP with plug-in estimators in the following examples (Refer to Appendix A for detailed derivations).

Illustration 4 (DML-IDP vs. Plug-in (PI) estimators for $P_{\mathbf{x}}(\mathbf{y})$ in Fig. (2a,2b,1)). *By Thm. 2, DML-IDP estimator for $P_{x_1,x_2}(y)$ in **Fig. (2a)** is consistent if estimates for either $\{P(v_i|pre(v_i))\}_{V_i\in\{X_1,X_2\}}$, or $\{P(v_i|pre(v_i))\}_{V_i\in\{X_1,Y\}}$, or $\{P(v_i|pre(v_i))\}_{V_i\in\{Z,Y\}}$ converge, while PI using Eq. (2) is consistent if estimates for $\{P(y|pre(y)), P(z|pre(z)), P(a|b,c), P(b|c), P(c)\}$ converge, where the variables are ordered as $C \prec B \prec A \prec X_1 \prec Z \prec X_2 \prec Y$.*

*DML-IDP estimator for $P_x(y_1,y_2,y_3,y_4,y_5)$ in **Fig. (2b)** is consistent if estimates $\{P(v_i|pre(v_i)\}_{V_i\in\{X,Y_1,Y_3,Y_4\}}$, or $\{P(v_i|pre(v_i)\}_{V_i\in\{X,Y_1,Y_5\}}$, or $\{P(v_i|pre(v_i)\}_{V_i\in\{Y_2,Y_5\}}$; and $\{P(v_i|pre(v_i))\}_{V_i\in\{C,Y_2\}}$, or $\{P(v_i|pre(v_i))\}_{V_i\in\{C,Y_3,Y_4\}}$, or $\{P(v_i|pre(v_i)\}_{V_i\in\{X,Y_1,Y_3,Y_4\}}$ converge, while PI using Eq. (6) is consistent if estimates for $\{P(v_i|pre(v_i))\}_{V_i\in\mathbf{V}}$ converge, where the order over $\mathbf{V}$ is $C \prec X \prec Y_1 \prec Y_2 \prec Y_3 \prec Y_4 \prec Y_5$.*

*DML-IDP for $P_x(y_1,y_2,y_3,y_4)$ in **Fig. (1)** is consistent if estimates for $P(x|r)$ or $\{P(y_2|x,r), P(y_3|y_2,x,r)\}$; and $\{P(v_i|pre(v_i))\}_{V_i\in\{R,X,Y_2,Y_3\}}$ or $P(y_4|pre(y_4))$ converge, while PI using Eq. (12) is consistent if estimates for $\{P(y_2|x,r), P(y_3|y_2,x,r), P(r), P(y_4|pre(y_4)), P(y_1)\}$ converge, where $Y_1 \prec R \prec X \prec Y_2 \prec Y_3 \prec Y_4$.*

6. Experiments

6.1. Experiments Setup

We evaluate DML-IDP for estimating $P_{\mathbf{x}}(\mathbf{y})$ in Fig. (2a,2b,1). We specify a SCM M for each PAG and generate datasets $\mathcal{D}$ from M. Details of the models and the data generating process are described in Appendix C. Throughout the experiments, the target causal effect is $\mu(\mathbf{x}) \equiv P_{\mathbf{x}}(\mathbf{Y} = 1)$, with ground-truth pre-computed. We compare DML-IDP with plug-in estimator (PI), the only available general-purpose estimator working for arbitrary causal functionals. Nuisance functions are estimated using standard techniques available in the literature (refer to Appendix C for details), e.g., conditional probabilities are

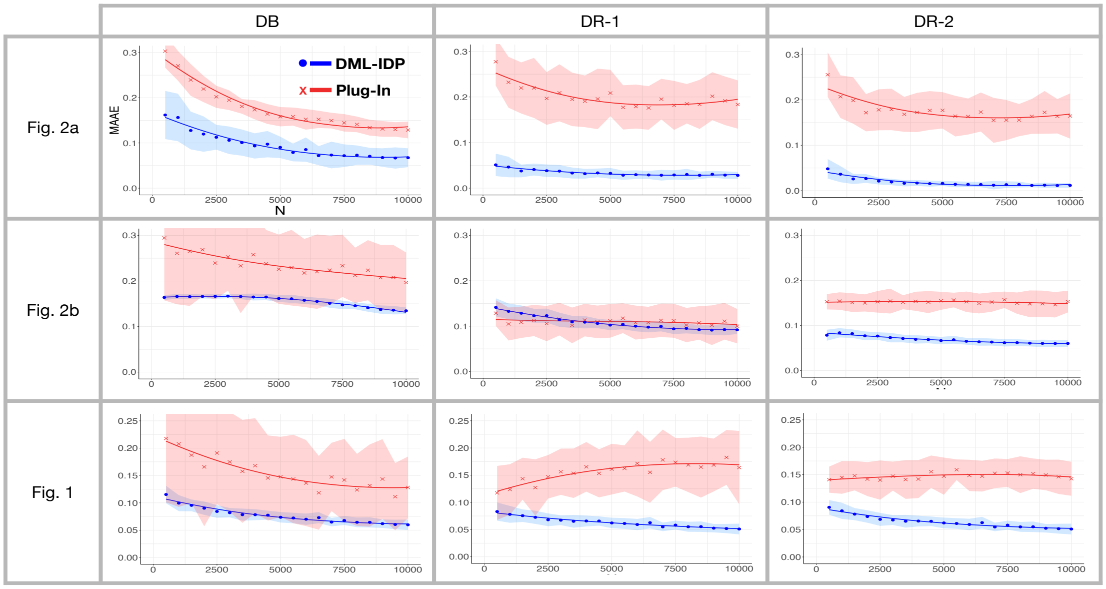

Figure 3: MAAE Plots for **(Top)** Fig. 2a, **(Middle)** Fig. 2b, and **(Bottom)** Fig. 1, under scenarios 'Debiasedness' ('DB') and 'Doubly Robustness' ('DR-1' and 'DR-2'). Shades represent one standard deviation.

estimated using a gradient boosting model XGBoost (Chen & Guestrin, 2016), which is known to be flexible.

Accuracy Measure Given a data set $\mathcal{D}$ with N samples, let $\widehat{\mu}_{\text{DML}}(\mathbf{x})$ and $\widehat{\mu}_{\text{PI}}(\mathbf{x})$ be the estimated $P_{\mathbf{x}}(\mathbf{Y} = 1)$ using DML-IDP and PI estimators. For each $\widehat{\mu} \in \{\widehat{\mu}_{\text{DML}}(\mathbf{x}), \widehat{\mu}_{\text{PI}}(\mathbf{x})\}$, we compute the average absolute error (AAE) as $|\mu(\mathbf{x}) - \hat{\mu}(\mathbf{x})|$ averaged over $\mathbf{x}$. We generate 100 datasets for each sample size N. We call the mean of the 100 AAEs the *mean average absolute error*, or MAAE, and its plot vs. the sample size N, the *MAAE plot*.

Simulation Strategy To show debiasedness ('DB') property, we add a 'converging noise' ϵ, decaying at a $N^{-\alpha}$ rate (i.e., $\epsilon \sim \text{Normal}(N^{-\alpha}, N^{-2\alpha})$) for $\alpha = 1/4$, to the estimated nuisance values to control the convergence rate of the estimators 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 et al., 2007).

6.2. Experimental Results

Debiasedness (DB) The MAAE plots for the debiasedness experiments for Fig. (2a,2b,1) are shown in the first column of Fig. 3. DML-IDP shows the debiasedness property against the converging noise decaying at $N^{-1/4}$ rates, while PI converges much slower for all three examples.

Doubly robustness (DR) The MAAE plots for the doubly robustness experiments are shown in the 2nd and 3rd columns of Fig. 3. Two misspecification scenarios are simulated for each example based on the results in Illustration 4. For Fig. 2a, nuisances $\{\widehat{P}(v_i|\text{pre}(v_i))\}$ for $V_i \in \{Y, Z\}$ in 'DR-1' and for $V_i \in \{Z, X_2\}$ in 'DR-2' are misspecified. For Fig. 2b, nuisances $\{\widehat{P}(v_i|\text{pre}(v_i))\}$ for $V_i \in \{Y_2, Y_5\}$ in 'DR-1' and for $V_i \in \{X, Y_1, Y_3, Y_4\}$ in 'DR-2' are misspecified. For Fig. 1, nuisances $\widehat{P}(y_4|\text{pre}(y_4))$ in 'DR-1' and $\{\widehat{P}(y_2|x,r), \widehat{P}(y_3|y_2,x,r)\}$ in 'DR-2' are misspecified. The results in all the scenarios support the doubly robustness of DML-IDP, whereas PI may fail to converge when misspecification is present.

7. Conclusion

We derived influence functions (Algo. 1, Thm. 1) and developed DML estimators named DML-IDP (Def. 5) for any causal effects identifiable given a Markov equivalence class of causal graphs represented as a PAG. DML-IDP estimators are guaranteed to have the property of debiasedness and doubly robustness (Thm. 2). Our experimental results demonstrate that these estimators are significantly more robust against model misspecification and slow convergence rate in learning nuisances compared to the only alternative estimator available in the literature, a plug-in estimator. We hope the new machinery developed here will allow more reliable and robust causal effect estimates by integrating modern ML methods that are capable of handling complex, high-dimensional data with causal learning and identification theory, paving the way towards a purely data-driven, end-to-end solution to causal effect estimation.

References

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

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

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

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

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

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

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

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

Fulcher, I. R., Shpitser, I., Marealle, S., and Tchetgen Tchetgen, E. J. Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):199–214, 2020.

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

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

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

Jaber, A., Zhang, J., and Bareinboim, E. A graphical criterion for effect identification in equivalence classes of causal diagrams. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018b.

Jaber, A., Zhang, J., and Bareinboim, E. Causal identification under markov equivalence: Completeness results. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2981–2989, 2019.

Jung, Y., Tian, J., and Bareinboim, E. Estimating causal effects using weighting-based estimators. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020a.

Jung, Y., Tian, J., and Bareinboim, E. Learning causal effects via weighted empirical risk minimization. *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, 2020b.

Jung, Y., Tian, J., and Bareinboim, E. Estimating identifiable causal effects through double machine learning. *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021.

Kallus, N. and Uehara, M. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. In *Proceedings of the 38th International Conference on Machine Learning*, 2020.

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

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

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

Lee, S. and Bareinboim, E. Causal effect identifiability under partial-observability. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Molina, J., Rotnitzky, A., Sued, M., and Robins, J. Multiple robustness in factorized likelihood models. *Biometrika*, 104(3):561–581, 2017.

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

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

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

Pearl, J. and Robins, J. Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Proceedings of the 11th Conference on Uncertainty in*

Artificial Intelligence, pp. 444–453. Morgan Kaufmann Publishers Inc., 1995.

Perkovic, E., Textor, J., Kalisch, M., and Maathuis, M. H. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *The Journal of Machine Learning Research*, 18 (1):8132–8193, 2017.

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

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

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

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

Rotnitzky, A. and Smucler, E. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188):1–86, 2020.

Rotnitzky, A., Robins, J., and Babino, L. On the multiply robust estimation of the mean of the g-functional. *arXiv preprint arXiv:1705.08582*, 2017.

Shpitser, I. and Pearl, J. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence*, pp. 1219, 2006.

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

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

Syrgkanis, V., Lei, V., Oprescu, M., Hei, M., Battocchi, K., and Lewis, G. Machine learning estimation of heterogeneous treatment effects with instruments. In *Proceedings of the 33th Annual Conference on Neural Information Processing Systems*, pp. 15193–15202, 2019.

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

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

Van der Laan, M. J. and Rose, S. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

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

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

Zadik, I., Mackey, L., and Syrgkanis, V. Orthogonal machine learning: Power and limitations. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5723–5731, 2018.

Zhang, J. *Causal inference and reasoning in causally insufficient systems*. PhD thesis, Citeseer, 2006.

Zhang, J. A characterization of markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, pp. 450–457, 2007.

Zhang, J. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9(Jul):1437–1474, 2008a.

Zhang, J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17): 1873–1896, 2008b.