Journal of Machine Learning Research 23 (2022) 1-76 Submitted 3/20; Revised 9/22; Published 10/22

Semiparametric Inference For Causal Effects In Graphical Models With Hidden Variables

Rohit Bhattacharya† RB17@WILLIAMS.EDU
Department of Computer Science
Williams College
Williamstown, MA 01267, USA

Razieh Nabi† RAZIEH.NABI@EMORY.EDU
Department of Biostatistics and Bioinformatics
Emory University
Atlanta, GA 30322, USA

Ilya Shpitser ILYAS@CS.JHU.EDU
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218, USA

† **Equal contribution**

Editor: Peter Spirtes

Abstract

Identification theory for causal effects in causal models associated with hidden variable directed acyclic graphs (DAGs) is well studied. However, the corresponding algorithms are underused due to the complexity of estimating the identifying functionals they output. In this work, we bridge the gap between identification and estimation of population-level causal effects involving a single treatment and a single outcome. We derive influence function based estimators that exhibit double robustness for the identified effects in a large class of hidden variable DAGs where the treatment satisfies a simple graphical criterion; this class includes models yielding the adjustment and front-door functionals as special cases. We also provide necessary and sufficient conditions under which the statistical model of a hidden variable DAG is nonparametrically saturated and implies no equality constraints on the observed data distribution. Further, we derive an important class of hidden variable DAGs that imply observed data distributions observationally equivalent (up to equality constraints) to fully observed DAGs. In these classes of DAGs, we derive estimators that achieve the semiparametric efficiency bounds for the target of interest where the treatment satisfies our graphical criterion. Finally, we provide a sound and complete identification algorithm that directly yields a weight based estimation strategy for any identifiable effect in hidden variable causal models.

Keywords: Unmeasured confounding; doubly robust estimation; nonparametric saturation; efficient influence function.

1. Introduction

Causal inference is concerned with the use of observed data to reason about cause-effect relationships encoded by counterfactual parameters, such as the population-level (or average) causal effect. Since counterfactual quantities are not directly observed in the data, they must be expressed as functionals of the observed data distribution using assumptions encoded in a causal model. The ease of conveying such assumptions pictorially via a directed acyclic graph (DAG) (Pearl, 2009; Spirtes et al., 2000) prompted further study of the identifiability of counterfactual quantities in causal models that factorize according to a DAG when some variables may be hidden or unobserved (Tian and Pearl, 2002a). This led to the development of a complete characterization of the identifiability of the average causal effect (ACE) of a given treatment on a given outcome in all hidden variable causal models associated with a DAG (Shpitser and Pearl, 2006; Huang and Valtorta, 2006).

Despite the sophistication of causal identification theory, estimators based on simple covariate adjustment remain the most common strategy for evaluating the ACE from data. Estimates obtained in this way are often biased due to the presence of unmeasured confounding and/or model misspecification. A popular approach for addressing the latter issue has been to use semiparametric estimators developed using the theory of influence functions (van der Vaart, 2000; Bang and Robins, 2005). The most popular of these estimators is known as the *augmented inverse probability weighted (AIPW)* estimator and is *doubly robust* in that it gives the analyst two chances to obtain a valid estimate for the ACE – either by specifying the correct model for the treatment assignment given observed covariates that render the treatment assignment ignorable, or by specifying the correct model for the dependence of the outcome on the treatment and these covariates. Recent work by Henckel et al. (2021) and Rotnitzky and Smucler (2020) yields methods for constructing statistically efficient versions of AIPW that take advantage of Markov restrictions implied on the obseved data by a fully observed causal model associated with a DAG.

If a causal model contains hidden variables, a.k.a. unmeasured confounders, causal inference becomes considerably more complicated. In the present work, we provide semiparametric estimators for the average causal effect of a single treatment variable on a single outcome variable in increasingly general scenarios, culminating in semiparametric estimators for *any* hidden variable causal model of a DAG in which this effect is identifiable. The front-door model (Pearl, 1995) is perhaps the simplest example of a graphical model with unmeasured confounders where no valid adjustment set exists, but the effect is still identifiable. An influence function based estimator has been derived for the identified functional in the front-door model by Fulcher et al. (2020). Weight-based estimators for a subclass of models considered in this paper, were studied in Jung et al. (2020). The authors in Jung et al. (2021) have a similar objective in bridging the gap between the identification and estimation theory in causal graphical models. Other related work includes numerical procedures for approximating the influence function proposed by Frangakis et al. (2015); Carone et al. (2019). However, such methods are either restricted to settings where simple covariate adjustment is valid or involve numerical approximations of the function itself which may be computationally prohibitive. There also exists a rich literature on semiparametric theory with instrumental variables; see for example Abadie (2003); Okui et al. (2012); Wang and Tchetgen Tchetgen (2018). However, these methods impose more assumptions, such

as monotonicity, exclusion restrictions, and causal relevance, than what is implied by the causal model itself represented via a DAG, possibly with hidden variables.

The paper is organized as follows. In Section 2, we provide a brief overview of causal graphical models. We first describe the causal and statistical models of DAGs where all variables are observed. We then move on to DAGs with hidden (or unmeasured) variables, and describe the latent projection of a hidden variable DAG into an acyclic directed mixed graph (ADMG). We discuss the district and topological factorizations of ADMGs, which are useful in deriving estimation results in Sections 4 and 5. Next, we introduce the general nested Markov factorization which is essential for our results in Section 6. We close this chapter with a more detailed overview of the results and their relation to semiparametric inference theory.

The nested Markov model of an ADMG encodes two types of equality restrictions: ordinary and generalized conditional independence constraints (a.k.a. Verma constrains). Such restrictions play an important role in deriving the tangent space of the model and the most efficient influence function based estimator for a given parameter. Hence, before diving into our estimation results for population-level effects, we take a closer look at these constraints in Section 3. We first provide a *sound* and *complete* procedure (Algorithm 1) for checking whether an ADMG imposes any equality restrictions on the observed data distribution, provided the hidden variables in the corresponding hidden variable DAGs are unrestricted. In the special case where the model is nonparametric saturated, i.e., no restrictions are imposed on the tangent space of the model, the influence function corresponding to the parameter of interest is unique. Thus the corresponding estimator is the most efficient in the given class of graphical models. We then define a class of ADMGs, termed *mb-shielded ADMGs*, for which the restrictions on the tangent space are all implied by simple conditional independence statements corresponding to a DAG model. Therefore, we can use known results on deriving the tangent space of such models (van der Vaart, 2000; Rotnitzky and Smucler, 2020), and derive estimators that achieve semiparametric efficiency bounds within this class. The results in this section are orthogonal to what the target of inference is.

In Section 4, we consider a class of ADMGs characterized via a simple graphical criterion for the treatment that we term *primal fixability*. In this class, the average causal effect of the treatment on any choice of the outcome is always identified. We provide two alternative representations for the identified functionals that directly yield two inverse probability weighting (IPW) type estimators. These representations, called *primal IPW* and *dual IPW*, use variationally independent components of the natural likelihood on the observed margin of the hidden variable DAG. We further derive the nonparametric influence function – the influence function in the nonparametric model – for the identified effect. The derivation is automated in the sense that in any ADMG where the treatment is primal fixable, the influence function can be mechanically derived by applying our results in Theorem 12. The resulting influence function based estimator can be viewed as an augmentation of the primal form. We call this *augmented primal IPW (APIPW)* and show that it is doubly robust in the two sets of models involved in the primal and dual IPW estimators. We close this chapter by describing a more stringent graphical criterion leading to identification via the *fixing* operation defined in Richardson et al. (2017). Causal effects identification via fixing can always be reformulated as covariate adjustment, and thus leads to estimation via the semiparametric augmented IPW (AIPW) estimator.

In Section 5, we focus on the class of mb-shielded ADMGs and discuss how we can exploit the constraints of such models to gain efficiency. Since mb-shielded ADMGs are equivalent to DAG models, we adapt known results for DAGs to obtain the space of all influence functions for the population-level effect of a primal fixable treatment on an outcome in this class of models. This space characterizes the *regular and asymptotically linear* estimators, which are $\sqrt{n}$-consistent and asymptotically normal, for our target parameter. We further derive the most efficient estimators within this class of causal models; i.e., influence function based estimators that achieve the *semiparametric efficiency bound*.

In Section 6, we describe semiparametric estimators for general classes of functionals representing identifiable causal effects of a single treatment on a single outcome, culminating with an estimator for *any* such functional. We propose the *nested IPW* estimator that generalizes IPW to all hidden variable causal models where the target parameter is identified. We propose a *sound* and *complete* algorithm (Algorithm 2) that derives the corresponding nested IPW estimator when possible. One of the interesting facts about this algorithm is that when the effect is identified, it outputs a functional that only relies on the conditional densities involving the variables in the district of the treatment; see next section for a description of district and other preliminaries.

In Section 7, we discuss alternative strategies for estimating the causal effect when treatment is not primal fixable. We illustrate the key results of this paper via a series of simulation analyses in Section 8, followed by conclusions in Section 9.

2. Overview of Causal Graphical Models

The cause-effect relationship between a single treatment T and an outcome Y is typically established through the use of potential outcomes, a.k.a. counterfactuals. For example, the potential outcomes $Y(1)$ and $Y(0)$ may be used to represent a hypothetical randomized controlled trial where units are randomly assigned to the treatment arm (corresponding to $T = 1$), or the control arm (corresponding to $T = 0$). The *average causal effect* (ACE) is a common target that is used to compare the distribution of such counterfactual random variables on the mean difference scale. That is, $\text{ACE} \equiv \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$. More generally, one could define a random variable $Y(t)$ corresponding to the potential outcome had treatment T been assigned to some specific value t. This allows for the contrast of arbitrary treatment assignments t and t' as $\mathbb{E}[Y(t)] - \mathbb{E}[Y(t')]$. Throughout the paper, we set our target of inference to be the mean of the counterfactual random variable $Y(t)$. That is,

$$\psi(t) \equiv \mathbb{E}[Y(t)]. \qquad \textit{(target parameter)} \tag{1}$$

2.1 Directed Acyclic Graphs (DAGs)

The target parameter $\psi(t)$ cannot be expressed as a function of the observed data, or in other words, is not identified, if no assumptions are made about the data generating process (Pearl, 2009). Causal graphs are a popular tool that can be used to provide an intuitive picture of substantive nonparametric assumptions made by the data analyst (Greenland et al., 1999; Richardson and Robins, 2013; Williams et al., 2018; Hünermund and Bareinboim, 2019).

A directed acyclic graph (DAG) $\mathcal{G}(V)$ is defined as a set of nodes V connected by directed edges such that there are no directed cycles. When the vertex set is clear from the given

context, we often abbreviate $\mathcal{G}(V)$ as simply $\mathcal{G}$. Causal models of a DAG $\mathcal{G}(V)$ are defined over counterfactual random variables $V_i(\text{pa}_i)$ for each $V_i \in V$, where $\text{pa}_{\mathcal{G}}(V_i)$ are the parents of V_i in $\mathcal{G}$ and pa_i is a set of values for $\text{pa}_{\mathcal{G}}(V_i)$. These counterfactuals can alternatively be viewed as being determined by a system of *structural equations* $f_i(\text{pa}_i, \epsilon_i)$ that map values pa_i, as well as values of an exogenous noise term ϵ_i to values of V_i (Pearl, 2009; Malinsky et al., 2019; Shpitser et al., 2020). Other counterfactuals may be defined from above via recursive substitution. Specifically, for any set $A \subseteq V$, and a variable V_i, we have:

$$V_i(a) \equiv V_i\big(a \cap \text{pa}_{\mathcal{G}}(V_i), \{V_j(a) : V_j \in \text{pa}_{\mathcal{G}}(V_i) \setminus A\}\big),$$

where $\{V_j(a) : V_j \in \text{pa}_{\mathcal{G}}(V_i) \setminus A\}$ is taken to mean the (recursively defined) set of counterfactuals associated with variables in $\text{pa}_{\mathcal{G}}(V_i) \setminus A$, had A been set to a.

For any set $A \subset V$, we denote the distribution of the potential outcomes $p(\{V_i(a) : V_i \in V \setminus A\})$ or $p(V(a))$ for short, where we assume for any $A_i \in A$, $A_i(a) = a_i$. In other words, the potential outcome $A_i(a)$ is the constant a_i, the value in a corresponding to A_i.

In a causal model of a DAG $\mathcal{G}$, $p(V(a))$ is identified by the g-formula functional (Robins, 1986):

$$p(V(a)) = \prod_{V_i \in V \setminus A} p(V_i \mid a \cap \text{pa}_{\mathcal{G}}(V_i), \text{pa}_{\mathcal{G}}(V_i) \setminus A). \qquad \textit{(g-formula)} \tag{2}$$

When A is the empty set, we obtain the familiar DAG factorization for $\mathcal{G}$ meaning that the causal model of a DAG $\mathcal{G}$ implies the statistical model of the DAG $\mathcal{G}$. That is, statistical models of a DAG $\mathcal{G}(V)$ are sets of distributions that factorize as,

$$p(V) = \prod_{V_i \in V} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)). \qquad \textit{(DAG factorization)} \tag{3}$$

Each missing edge between pairs of variables in a DAG $\mathcal{G}$ imply conditional independences in $p(V)$. These can be read off directly from $\mathcal{G}$ via the well-known d-separation criterion (Pearl, 2009). That is, for disjoint sets X, Y, and Z, the following *global Markov property* holds: $(X \perp\!\!\!\perp_{\text{d-sep}} Y \mid Z)_{\mathcal{G}} \implies (X \perp\!\!\!\perp Y \mid Z)_{p(V)}$. When the context is clear, we simply use $X \perp\!\!\!\perp Y \mid Z$ to denote conditional independence between X and Y given Z.

In all causal models of a DAG $\mathcal{G}$, the target parameter $\psi(t)$ is identified via the back-door adjustment formula as follows,

$$\psi(t) = \mathbb{E}\big[\mathbb{E}[Y \mid T = t, \text{pa}_{\mathcal{G}}(T)]\big]. \qquad \textit{(adjustment functional)} \tag{4}$$

Once the target parameter is identified, causal inference reduces to an estimation problem of the identifying functional. There exist several estimators for the adjustment functional, such as plug-in, inverse probability weighting (IPW), and augmented inverse probability weighting (AIPW) (Robins et al., 1994; Hahn, 1998; Robins, 2000; van der Laan and Rose, 2011; Kennedy et al., 2017). An overview of these estimators can be found in Appendix D.1.

2.2 DAGs with Hidden Variables

While estimation theory for fully observed causal models represented by DAGs is well developed, causal models most relevant to practical applications are sure to contain variables

that are unmeasured or hidden to the data analyst. In such cases, the observed data distribution $p(V)$ can be viewed as a margin of a distribution $p(V \cup H)$ associated with a DAG $\mathcal{G}(V \cup H)$ where vertices in V correspond to observed variables and vertices in H correspond to unmeasured or hidden variables. Two complications arise from the presence of hidden variables. First, the target parameter $\psi(t)$ may not always be identified as a function of the observed data law, and second, parameterizations of latent variable models are generally not globally identified and may contain singularities (Drton, 2009).

A natural alternative to the latent variable model is one that places no restrictions on $p(V)$ aside from those implied by the Markov restrictions given by the factorization of $p(V \cup H)$ with respect to $\mathcal{G}(V \cup H)$. It was shown by Evans (2018) that all equality constraints implied by such a factorization are captured by a nested factorization of $p(V)$ with respect to an *acyclic directed mixed graph (ADMG)* $\mathcal{G}(V)$ derived from $\mathcal{G}(V \cup H)$ via the latent projection operation described by Verma and Pearl (1990). Such an ADMG is a smooth supermodel of infinitely many hidden variable DAGs that share the same identification theory for $\psi(t)$, and imply the same equality constraints on the margin $p(V)$ (Richardson et al., 2017; Evans and Richardson, 2019). Thus, our use of ADMGs for identification and estimation of the target $\psi(t)$ is without loss of generality.

The latent projection of a hidden variable DAG $\mathcal{G}(V \cup H)$ onto the observed variables V is an ADMG $\mathcal{G}(V)$ with directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges constructed as follows. The edge $V_i \rightarrow V_j$ exists in $\mathcal{G}(V)$ if there exists a directed path from V_i to V_j in $\mathcal{G}(V \cup H)$ with all intermediate vertices in H. An edge $V_i \leftrightarrow V_j$ exists in $\mathcal{G}(V)$ if there exists a collider-free path (i.e., there are no consecutive edges of the form $\rightarrow \circ \leftarrow$) from V_i to V_j in $\mathcal{G}(V \cup H)$ with all intermediate vertices in H, such that the first edge on the path is an incoming edge into V_i and the final edge is an incoming edge into V_j. An example of latent projection is provided in Appendix B. Conditional independences in the observed distribution $p(V)$ can be read off from the ADMG $\mathcal{G}(V)$ by a simple analogue of the d-separation criterion, known as m-separation, that generalizes the notion of a collider to include mixed edges of the form $\rightarrow \circ \leftrightarrow$, $\leftrightarrow \circ \leftarrow$, and $\leftrightarrow \circ \leftrightarrow$, (Richardson, 2003).

The bidirected connected components of an ADMG $\mathcal{G}(V)$ partition its vertices into distinct subsets known as *districts*. A set $S \subseteq V$ is a district in $\mathcal{G}(V)$ if it forms a maximal connected component via only bidirected edges. We use $\text{dis}_{\mathcal{G}}(V_i)$ to denote the district of V_i in $\mathcal{G}$, which includes V_i itself, and $\mathcal{D}(\mathcal{G})$ to denote the set of all districts in $\mathcal{G}$.

2.2.1 District and Topological Factorization of ADMGs

We first define a simple factorization of $p(V)$ relative to an ADMG in terms of its districts and objects known as kernels. A kernel $q_V(V \mid W)$ is a mapping from values of W to normalized densities over V. That is, $\sum_V q_V(V \mid W = w) = 1, \forall w \in W$ (Lauritzen, 1996). For any set of variables $X \subseteq V$, marginalization and conditioning in a kernel are defined in the usual way, i.e., $q_{V \setminus X}(V \setminus X \mid W) \equiv \sum_X q_V(V \mid W)$ and $q_V(V \setminus X \mid X, W) \equiv \frac{q_V(V|W)}{q_V(X|W)}$.

A distribution $p(V)$ is said to *district factorize* or *Tian factorize* with respect to an ADMG $\mathcal{G}(V)$ if

$$p(V) = \prod_{D \in \mathcal{D}(\mathcal{G})} q_D(D \mid \text{pa}_{\mathcal{G}}(D)), \qquad \textit{(District ADMG factorization)} \tag{5}$$

where the parents of a set of vertices D is defined as the set of parents of D not already in D, i.e., $\mathrm{pa}_{\mathcal{G}}(D) \equiv \left(\bigcup_{D_i \in D} \mathrm{pa}_{\mathcal{G}}(D_i)\right) \setminus D$. We follow the same convention for children of a set S, denoted $\mathrm{ch}_{\mathcal{G}}(S)$. For other standard genealogical relations defined for a single vertex V_i, such as ancestors $\mathrm{an}_{\mathcal{G}}(V_i) \equiv \{V_j \in V \mid \exists\ V_j \to \cdots \to V_i \text{ in } \mathcal{G}\}$ and descendants $\mathrm{de}_{\mathcal{G}}(V_i) \equiv \{V_j \in V \mid \exists\ V_i \to \cdots \to V_j \text{ in } \mathcal{G}\}$ both of which include V_i itself by convention, the extension to a set S uses the disjunctive definition which also includes the set itself. For example, $\mathrm{an}_{\mathcal{G}}(S) = \bigcup_{S_i \in S} \mathrm{an}_{\mathcal{G}}(S_i)$. A list of notation and definitions used in this paper can be found in Appendix A.

The use of q in place of p in Eq. 5 emphasizes the fact that these factors are not necessarily ordinary conditional distributions. Each factor $q_D(D \mid \mathrm{pa}_{\mathcal{G}}(D))$ can in fact be treated as a post-intervention distribution where all variables outside of D are intervened on and held fixed to some constant value (Tian and Pearl, 2002a). Hence, we use $q_S(\cdot \mid \cdot)$ to denote a kernel where only variables in S are random and all others are fixed.

Tian and Pearl (2002a) showed that each kernel $q_D(D \mid \mathrm{pa}_{\mathcal{G}}(D))$ appearing in Eq. 5 is a function of $p(V)$ as follows. Define the *Markov blanket* of a vertex V_i as the district of V_i and the parents of its district, excluding V_i itself. That is, $\mathrm{mb}_{\mathcal{G}}(V_i) \equiv (\mathrm{dis}_{\mathcal{G}}(V_i) \cup \mathrm{pa}_{\mathcal{G}}(\mathrm{dis}_{\mathcal{G}}(V_i))) \setminus V_i$. Consider a valid topological order τ on all k vertices in V, that is, a sequence $(V_1, \ldots, V_k)$ such that no vertex appearing later in the sequence is an ancestor of vertices earlier in the sequence. Let $\{\preceq_\tau V_i\}$ denote the set of vertices that precede V_i in this sequence, including V_i itself. Define the *Markov pillow* of V_i, denoted by $\mathrm{mp}_{\mathcal{G}}(V_i)$, as its Markov blanket in a subgraph restricted to V_i and its predecessors according to the topological ordering $\prec_\tau$. We suppress the dependence of $\mathrm{mp}_{\mathcal{G}}(V_i)$ on $\prec_\tau$ for notational conciseness. More formally, $\mathrm{mp}_{\mathcal{G}} \equiv \mathrm{mb}_{\mathcal{G}_S}(V_i)$ where $S = \{\preceq_\tau V_i\}$, and $\mathcal{G}_S$ is the subgraph of $\mathcal{G}$ that is restricted to vertices in S and the edges between these vertices. Then for each $D \in \mathcal{D}(\mathcal{G})$,

$$q_D(D \mid \mathrm{pa}_{\mathcal{G}}(D)) = \prod_{D_i \in D} p(D_i \mid \mathrm{mp}_{\mathcal{G}}(D_i)). \qquad \textit{(Identification of district factors)} \tag{6}$$

This leads to a factorization of the observed law as a product of simple conditional factors according to the given valid topological order,

$$p(V) = \prod_{V_i \in V} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)). \qquad \textit{(Topological ADMG factorization)} \tag{7}$$

The above factorization (and the district factorization from which it is derived) does not always capture every equality restriction in $p(V)$ implied by the Markov property of the underlying hidden variable DAG $\mathcal{G}(V \cup H)$. However, it is particularly simple to work with, and under some conditions, which we derive in Section 3, is capable of capturing all such restrictions.

2.2.2 Nested Markov Factorization of ADMGs

We first introduce some graphical concepts used to describe the *nested Markov factorization* of an ADMG that captures all equality constraints on the observed margin $p(V)$. This factorization is defined on conditional ADMGs and kernels derived from $\mathcal{G}(V)$ and $p(V)$ via a fixing operation. *Conditional ADMGs* (CADMGs) $\mathcal{G}(V, W)$ are ADMGs whose vertices

can be partitioned into random variables V and fixed variables W, with the restriction that only outgoing edges may be adjacent to variables in W (Richardson et al., 2017). CADMGs are often used to represent post-intervention distributions where variables in W have been intervened on or *fixed*. For any random variable $V_i \in V$ in a CADMG $\mathcal{G}(V, W)$, the usual definitions of genealogical relations and other special sets, such as parents, descendants, Markov blankets, and Markov pillows, extend naturally by allowing for the inclusion of fixed variables into these sets. In a CADMG $\mathcal{G}(V, W)$, districts are only defined for elements of V.

A vertex $V_i \in V$ is said to be *fixable* in $\mathcal{G}(V, W)$ if $\text{dis}_{\mathcal{G}}(V_i) \cap \text{de}_{\mathcal{G}}(V_i) = \{V_i\}$. In words, V_i is fixable if there are no bidirected paths from V_i to any of its descendants. The graphical operation of fixing V_i, denoted by $\phi_{V_i}(\mathcal{G})$, yields a new CADMG $\mathcal{G}(V \setminus V_i, W \cup V_i)$ where bidirected and directed edges into V_i are removed, and V_i is fixed to a particular value v_i. Given a kernel $q_V(V \mid W)$ associated with the CADMG $\mathcal{G}(V, W)$, the corresponding probabilistic operation of fixing, denoted by $\phi_{V_i}(q_V; \mathcal{G})$, yields a new kernel

$$\phi_{V_i}(q_V; \mathcal{G}) \equiv q_{V \setminus V_i}(V \setminus V_i \mid W \cup V_i) \equiv \frac{q_V(V \mid W)}{q_V(V_i \mid \text{mb}_{\mathcal{G}}(V_i), W)}. \quad \textit{(Probabilistic fixing operator)} \tag{8}$$

The definition of fixability can be extended to a set of vertices S by requiring that there exists an ordering $(S_1, \ldots, S_p)$ such that S_1 is fixable in $\mathcal{G}$, S_2 is fixable in $\phi_{S_1}(\mathcal{G})$, and so on. Such an ordering is said to form a valid fixing sequence for S. It is known that any two valid fixing sequences on S yield the same CADMG, which we will denote by $\phi_S(\mathcal{G}(V, W))$. Fix a CADMG $\mathcal{G}(V, W)$ and a corresponding kernel $q(V \mid W)$. Given a valid fixing sequence σ_S on $S \subseteq V$ valid in $\mathcal{G}(V, W)$, define $\phi_{\sigma_S}(q_V; \mathcal{G})$ inductively to be $q(V \mid W)$ when S is empty, and $\phi_{\sigma_S \setminus S_1}(\phi_{S_1}(q_V; \mathcal{G}); \phi_{S_1}(\mathcal{G}))$ otherwise, where $\sigma_S \setminus S_1$ corresponds to the remainder of the sequence after removing the first element S_1. A concrete example demonstrating sequential applications of the graphical and probabilistic operations of fixing can be found in Appendix C.

A set D is called *intrinsic* in $\mathcal{G}(V)$ if $V \setminus D$ is fixable in $\mathcal{G}(V)$ and $\phi_{V \setminus D}(\mathcal{G}(V))$ contains a single district. A distribution $p(V)$ is said to satisfy the nested Markov factorization relative to an ADMG $\mathcal{G}(V)$ if there exists a set of kernels $q_D(D \mid \text{pa}_{\mathcal{G}}(D))$, one for every D intrinsic in $\mathcal{G}(V)$, such that for every fixable set S and every valid fixing sequence σ_S,

$$\phi_{\sigma_S}(p(V); \mathcal{G}) = \prod_{D \in \mathcal{D}(\phi_S(\mathcal{G}))} q_D(D \mid \text{pa}_{\mathcal{G}}(D)). \quad \textit{(Nested Markov factorization)} \tag{9}$$

In words, the nested Markov factorization states that every kernel that can be derived via a valid sequence of fixing satisfies the district factorization with respect to the CADMG obtained by this sequence, and each of the kernels appearing in the factorization corresponds to intrinsic sets. Given a distribution that satisfies the nested Markov factorization, for any fixable set S, applying any two distinct valid sequences σ_S^1, σ_S^2 to $p(V)$ and $\mathcal{G}(V)$ also yields the same kernel, which we can then denote as $\phi_S(p(V); \mathcal{G}(V))$ (Richardson et al., 2017).

2.3 Brief Overview of Semiparametric Inference

Assume a semiparametric model $\mathcal{M} = \{p(Z; \eta) : \eta \in \Gamma\}$ where Γ is the parameter space and η is the parameter indexing a specific distribution. The *tangent space* of a statistical model

$\mathcal{M}$ is defined as the mean-square closure of all linear combinations of scores in corresponding parametric submodels for $\mathcal{M}$. It is well-known that the tangent space in the statistical model of a DAG $\mathcal{G}(V)$, denoted by Λ, can be partitioned into a direct sum of orthogonal subspaces (Bickel et al., 1993; van der Vaart, 2000; Tsiatis, 2007). That is,

$$\Lambda \equiv \oplus_{V_i \in V} \Lambda_i, \qquad \textit{(Tangent space of statistical models of DAGs)} \tag{10}$$

where $\Lambda_i \equiv \big\{\alpha_i(V_i, \text{pa}_{\mathcal{G}}(V_i)) \in \mathbb{H} \mid \mathbb{E}[\alpha_i \mid \text{pa}_{\mathcal{G}}(V_i)] = 0\big\}$, and $\mathbb{H}$ denotes the *Hilbert space* defined as the space of all mean-zero scalar functions, equipped with the inner product $\mathbb{E}[h_1 \times h_2], \forall h_1, h_2 \in \mathbb{H}$. If $\mathcal{G}$ is a complete DAG, i.e., every vertex is connected to every other vertex, then there exist no independence relations between any sets of variables. In such scenarios, the tangent space equals the entire Hilbert space. In general, any statistical model with tangent space Λ, where $\Lambda = \mathbb{H}$, is said to be *nonparametric saturated* (NPS).

We are often interested in a function $\psi : \eta \in \Gamma \mapsto \psi(\eta) \in \mathbb{R}$; i.e., a parameter that maps the distribution P_η to a scalar number in $\mathbb{R}$, such as an identified average causal effect. (For brevity, we sometimes use ψ instead of $\psi(\eta)$, which should be obvious from context.)

An estimator $\widehat{\psi}_n$ of a scalar parameter ψ based on n i.i.d copies $Z_1, \ldots, Z_n$ drawn from $p(Z; \eta)$, is *asymptotically linear* if there exists a measurable random function $U_\psi(Z)$ with mean zero and finite variance such that

$$\sqrt{n} \times (\widehat{\psi}_n - \psi) = \frac{1}{\sqrt{n}} \times \sum_{i=1}^{n} U_\psi(Z_i) + o_p(1),$$

where $o_p(1)$ is a term that converges in probability to zero as n goes to infinity. The random variable $U_\psi(Z)$ is called the *influence function* (IF) of the estimator $\widehat{\psi}_n$. The analysis is often restricted to regular and asymptotically linear (RAL) estimators to exclude super efficient estimators, such as the Hodges' estimator, whose behavior is difficult to analyze in some parts of the model space. The RAL estimator $\widehat{\psi}_n$ is *$\sqrt{n}$-consistent and asymptotically normal* (CAN), with asymptotic variance equal to the variance of its influence function U_ψ,

$$\sqrt{n} \times (\widehat{\psi}_n - \psi) \xrightarrow{d} N\big(0,\ \text{var}(U_\psi)\big).$$

Influence functions in semiparametric models are derived as normalized elements of the orthogonal complement of the tangent space of the model. The orthogonal complement of the tangent space is defined as $\Lambda^\perp = \{h \in \mathbb{H} \mid \mathbb{E}[h \times h'] = 0, \forall h' \in \Lambda\}$; $\mathbb{H} = \Lambda \oplus \Lambda^\perp$, where $\oplus$ denotes the direct sum, and $\Lambda \cap \Lambda^\perp = \{0\}$. The vector space $\Lambda^\perp$ is of particular importance because we can construct the class of all influence functions, denoted by $\mathcal{U}$, as $\mathcal{U} = \{U_\psi\} + \Lambda^\perp$. In other words, upon knowing a single influence function U_ψ and $\Lambda^\perp$, we can obtain the class of all possible RAL estimators that admit the CAN property. Out of all IFs in $\mathcal{U}$, there exists a unique one which lies in the tangent space Λ and yields the most efficient RAL estimator by recovering the *semiparametric efficiency bound*. This efficient influence function can be obtained by projecting any influence function, call it U^*_ψ, onto the tangent space Λ. This operation is denoted by $U^{\text{eff}}_\psi = \pi[U^*_\psi \mid \Lambda]$, where U^{eff}_ψ denotes the efficient influence function. In a nonparametric saturated model (one with an unrestricted tangent space), the IF is unique and the corresponding estimator is the one that achieves the semiparametric efficiency bound. For a more detailed description of the concepts outlined here, see Appendix D and (van der Vaart, 2000; Tsiatis, 2007).

3. A Class of ADMGs Observationally Equivalent to DAGs

Efficient semiparametric estimators must take advantage of constraints that restrict the tangent space of the model. The nested Markov model of an ADMG encodes two types of equality constraints: ordinary conditional independences and generalized conditional independences a.k.a Verma constraints (Verma and Pearl, 1990). For an example of the latter constraint consider the ADMG shown in Fig. 1(a). The m-separation criterion can be used to show that the absence of an edge between T and Y does not correspond to any ordinary conditional independence between these variables. However, the nested Markov factorization in Eq. 9 implies that the kernel derived by fixing all variables except Y (following any valid fixing sequence, e.g., (L, M, T)) district factorizes with respect to the CADMG shown in Fig. 1(b). That is, given a distribution $p(V)$ satisfying the nested Markov factorization with respect to $\mathcal{G}(V)$ in Fig. 1(a) we have,

$$\phi_{\{L,M,T\}}(p(V); \mathcal{G}) = \sum_{M} p(M \mid T) \times p(Y \mid T, M, L) = q_Y(Y \mid L). \tag{11}$$

The first equality follows from the definition of the fixing operator; the second follows from the nested Markov factorization. Eq. 11 implies that $\sum_M p(M \mid T) \times p(Y \mid T, M, L)$ is not a function of T; this corresponds to a Verma constraint in $\mathcal{G}$.

It is not easy to see how Verma-type restrictions can be translated into efficiency gains in estimators in general. On the other hand, conditional independence restrictions that form Markov models associated with DAGs make efficient estimators easier to derive. This motivates our results in this section. First, in Section 3.1, we provide a sound and complete algorithm that characterizes when the nested Markov model of an ADMG $\mathcal{G}(V)$ is nonparametric saturated (NPS), meaning that the model imposes no equality restrictions (ordinary or generalized) on $p(V)$. In Section 3.2 we describe a class of ADMGs, called *mb-shielded ADMGs*, that are observationally equivalent to DAGs. In other words, when an ADMG is mb-shielded, all equality constraints in the model are implied by ordinary conditional independences according to a valid topological order. These results will allow us, in Section 5, to derive efficient semiparametric estimators for certain identified counterfactual mean parameters by examining the form of the tangent space of the ADMG models in the classes we describe. However, the results in this section are general, and can be applied to any target parameter.

3.1 Algorithm to Detect Nonparametric Saturation

The model implied by a complete DAG (a complete graph is one where all vertices are pairwise connected by a directed or bidirected edge) is nonparametric saturated, since the set of constraints from the local Markov property of the DAG model – which states that each variable is independent of its non-descendant non-parents given its parents, and is a small list that implies all other constraints in DAG models – is empty. Further, any model corresponding to a DAG that is not complete is not saturated – a missing edge in a DAG can always be translated into a statement in the local Markov property of a DAG model. By contrast, an ADMG that is not complete may still represent a nonparametric saturated nested Markov model. Consider a modification to the ADMG in Fig. 1(a) where we add the bidirected edge $L \leftrightarrow Y$; the resulting graph is shown in Fig. 1(c). Though this new ADMG

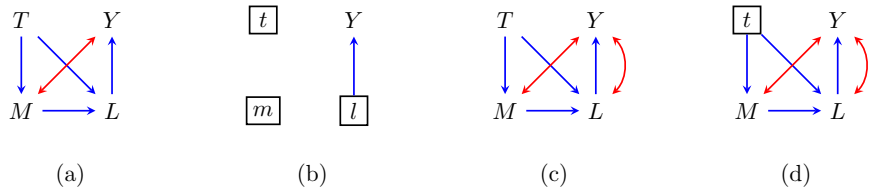


Figure 1: (a) Example of an ADMG where the missing edge betweem T and Y corresponds to a Verma constraint; (b) CADMG corresponding to the kernel $q_Y(T \mid L)$ that shows the Verma constraint in (a); (c) Example of an ADMG where the missing edge between T and Y does not correspond to any equality constraint; (d) The CADMG $\phi_{\neg\{Y\}}(\mathcal{G})$ when checking for equality constraints in (c).

still lacks an edge between T and Y, there is no longer any Verma constraint implied by the nested Markov factorization. In fact, it can be shown that the model is NPS.

Algorithm 1 provides a description of our procedure for checking if the nested Markov model of an ADMG $\mathcal{G}(V)$ is nonparametric saturated. Line 4 uses the notation $\phi_{\neg S}(\mathcal{G})$ to denote the (unique) CADMG obtained by recursively fixing as many vertices as possible in the set $V \setminus S$. As examples of this notation, the CADMG in Fig. 1(b) shows $\phi_{\neg\{Y\}}(\mathcal{G})$ when the original ADMG $\mathcal{G}$ is the one shown in Fig. 1(a), and the CADMG in Fig. 1(d) shows $\phi_{\neg\{Y\}}(\mathcal{G})$ for the ADMG in Fig. 1(c).

The intuition behind the algorithm is as follows. Rather than directly examining the nested Markov model of an arbitrary ADMG $\mathcal{G}$, we examine the model of its maximal arid projection (Shpitser et al., 2018) $\mathcal{G}^a$. For our purposes, $\mathcal{G}^a$ has the desirable property that its model is observationally equivalent to the nested Markov model of $\mathcal{G}$. In addition, we show (as part of the soundness proof for Algorithm 1) that maximal arid graphs have a property similar to DAGs wrt equality constraints: any missing edge in a maximal arid graph $\mathcal{G}^a$ corresponds to an (ordinary or generalized) equality constraint in the model. Algorithm 1 is designed around these facts. In particular, it returns "NPS" when the maximal arid projection $\mathcal{G}^a$ of the input ADMG $\mathcal{G}$ is a complete graph (one where all vertices are pairwise connected.) Since $\mathcal{G}$ and $\mathcal{G}^a$ imply the same nested Markov model (Shpitser et al., 2018), the nested Markov model of $\mathcal{G}$ is nonparametrically saturated by Corollary 17 described in the Appendix[1]. The algorithm returns "not NPS" when the maximal arid projection $\mathcal{G}^a$ has no edge (directed or bidirected) between at least one pair of vertices, in particular the pair (V_i, V_j) for which the checks in line 4 succeeded.

Algorithm 1 is also computationally tractable; it runs in polynomial time with respect to the number of vertices and edges in the graph $\mathcal{G}$. The complexity of the outer and inner loops is $\mathcal{O}(|V|^2)$. Naive implementations for computing CADMGs such as $\phi_{\neg\{V_i\}}$ and $\phi_{\neg\{V_i,V_j\}}$ also have polynomial complexity $\mathcal{O}(|V|^2 + |V| \times |E|)$ as it involves repeated applications of depth first search (popular algorithms for which have linear complexity $\mathcal{O}(|V| + |E|)$ (Tarjan, 1972)) in order to determine the fixability of a set of vertices.

1. Briefly, any complete ADMG is mb-shielded (Theorem 2), and hence equivalent to a complete DAG.

Algorithm 1 Check Nonparametric Saturation ($\mathcal{G}$)

1: Let τ be a valid topological order for V
2: **for** each distinct (V_i, V_j) pair in V **do**
3: Assume wlog $V_j \prec_\tau V_i$ and let D be the district of V_i in $\phi_{\neg\{V_i\}}(\mathcal{G})$
4: **if** $V_j \notin \text{pa}_{\mathcal{G}}(D_i)$ for all $D_i \in D$ **and** $\phi_{\neg\{V_i,V_j\}}(\mathcal{G})$ has more than one district **then**
5: **return** not NPS
6: **return** NPS

The following theorem formalizes the soundness and completeness properties of our algorithm. That is, Algorithm 1 correctly declares the model to be NPS if it is indeed NPS; when the model is declared as not NPS, the form of an equality constraint is provided.

Theorem 1 (Soundness and completeness of Algorithm 1)
Algorithm 1 is sound and complete for determining the absence of equality constraints in the nested Markov model of an ADMG $\mathcal{G}(V)$.

3.1.1 Example: Nonparametric Saturation

As an example, we demonstrate applying Algorithm 1 to the ADMGs in Fig. 1(a) and (c). As all pairs of vertices besides T and Y are connected via a directed or bidirected edge in these ADMGs, the conditions in line 4 trivially evaluate to False for these pairs; we thus focus on steps executed when examining the pair (T, Y). Since T is an ancestor of Y we have $T \prec_\tau Y$. The algorithm first examines the CADMG $\phi_{\neg\{Y\}}(\mathcal{G})$. In the case of Fig. 1(a), this CADMG corresponds to the one shown in Fig. 1(b), and we see T is not a parent of any member of the district of Y (which is just Y in this case.) The CADMG $\phi_{\neg\{T,Y\}}(\mathcal{G})$ is similar to the one shown in Fig. 1(b) except T remains a random vertex. In this CADMG, there are two distinct districts $\{T\}$ and $\{Y\}$. Hence both conditions in line 4 are met and the algorithm returns that the model is not NPS as expected. When we apply the algorithm to the ADMG shown in Fig. 1(c) we obtain the CADMG $\phi_{\neg\{Y\}}(\mathcal{G})$ shown in Fig. 1(d). In this case T is a parent of M and L which are both in the district of Y and so the algorithm returns that the model is NPS, once again matching the discussion at the beginning of this section.

3.2 Mb-shielded ADMGs

In this section we describe a large class of ADMGs where all equality constraints are implied by ordinary conditional independences according to a valid topological order (resembling the local Markov property for fully observed DAG models) and derive the tangent space of such ADMGs. As mentioned earlier, deriving efficient estimators in such models is considerably easier; the derivation of the tangent space for arbitrary ADMGs is a challenging problem left for future work.

First, assume the existence of a class of ADMGs where, given a topological order τ, all equality constraints implied by the ADMG $\mathcal{G}(V)$ can be written as ordinary conditional independence statements of the form,

$$V_i \perp\!\!\!\perp \{\prec_\tau V_i\} \setminus \text{mp}_{\mathcal{G}}(V_i) \mid \text{mp}_{\mathcal{G}}(V_i). \tag{12}$$

Such a property immediately implies that the topological factorization of the observed data distribution $p(V)$ shown in Eq. 7 captures all equality constraints implied by the ADMG $\mathcal{G}(V)$. A sound criterion for identifying ADMGs that satisfy this property is to check that an edge between two vertices V_i and V_j in $\mathcal{G}$ is absent only if $V_i \notin \text{mb}_{\mathcal{G}}(V_j)$ *and* $V_j \notin \text{mb}_{\mathcal{G}}(V_i)$. We call this class of ADMGs *mb-shielded ADMGs*, as pairs of vertices are always adjacent if either one is in the Markov blanket of the other. We formalize this criterion in the following theorem, and show that all equality constraints in mb-shielded ADMGs are implied by the set of ordinary conditional independence statements in Eq. 12.

Theorem 2 (mb-shielded ADMGs)
Consider a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where an edge between two vertices is absent only if $V_i \notin \text{mb}_{\mathcal{G}}(V_j)$ and *$V_j \notin \text{mb}_{\mathcal{G}}(V_i)$. Then, given any valid topological order on V, all equality constraints in $p(V)$ are implied by the set of restrictions: $V_i \perp\!\!\!\perp \{\prec V_i\} \setminus \text{mp}_{\mathcal{G}}(V_i) \mid \text{mp}_{\mathcal{G}}(V_i)$, $\forall V_i \in V$.*

According to Theorem 2 and the local Markov property of DAGs, we can see that all the equality constraints in an mb-shielded ADMG are DAG-like. One of the implications of Theorem 2 is that the tangent space in an mb-shielded ADMG is identical to the one of a DAG provided in display (10) by replacing each $\text{pa}_{\mathcal{G}}(V_i)$ with $\text{mp}_{\mathcal{G}}(V_i)$. This follows directly from Lemma 1.6 in van der Laan and Robins (2003), but we reiterate these results below for the sake of completeness.

Lemma 3 (Λ and $\Lambda^{\perp}$ in mb-shielded ADMGs)
Consider the statistical model $\mathcal{M}(\mathcal{G})$ where $\mathcal{G}(V)$ is an mb-shielded ADMG. The tangent space of $\mathcal{M}(\mathcal{G})$ is given by a direct sum of mutually orthogonal spaces: $\Lambda = \oplus_{V_i \in V} \Lambda_i$, where

$$\begin{aligned}\Lambda_i &= \big\{\alpha_i(V_i, \text{mp}_{\mathcal{G}}(V_i)) \in \mathbb{H} \text{ s.t. } \mathbb{E}[\alpha_i \mid \text{mp}_{\mathcal{G}}(V_i)] = 0\big\} \\ &= \big\{\alpha_i(V_i, \text{mp}_{\mathcal{G}}(V_i)) - \mathbb{E}[\alpha_i \mid \text{mp}_{\mathcal{G}}(V_i)],\ \forall \alpha_i(V_i, \text{mp}_{\mathcal{G}}(V_i)) \in \mathbb{H}\big\}.\end{aligned}$$

In addition, the projection of an element $h(V) \in \mathbb{H}$ onto Λ_i, denoted by h_i, is given by $h_i \equiv \Pi[h(V) \mid \Lambda_i] = \mathbb{E}\big[h(V) \mid V_i, \text{mp}_{\mathcal{G}}(V_i)\big] - \mathbb{E}\big[h(V) \mid \text{mp}_{\mathcal{G}}(V_i)\big]$. Consequently, the orthogonal complement of the tangent space $\Lambda^{\perp}$ is given as follows,

$$\Lambda^{\perp} = \Big\{ \sum_{V_i \in V} \alpha_i(V_1, \ldots, V_i) - \mathbb{E}\Big[\alpha_i(V_1, \ldots, V_i) \ \Big|\ V_i, \text{mp}_{\mathcal{G}}(V_i)\Big] \Big\},$$

where $\alpha_i(V_1, \ldots, V_i)$ is any function of V_1 through V_i in $\mathbb{H}$, such that $\mathbb{E}[\alpha_i \mid V_1, \ldots, V_{i-1}] = 0$.

In the following section, we provide new IPW and influence function based estimators for our counterfactual mean target $\psi(t)$, in ADMGs that satisfy a simple graphical criterion that we term *primal fixability*. In Section 5, we derive estimators that achieve the semiparametric efficiency bound in mb-shielded ADMGs that satisfy primal fixability.

4. Average Causal Effects: Primal Fixability of Treatment T

Unmeasured confounding (bidirected arrows in ADMGs) complicates identification, and hence, estimation of causal effects. Consider the ADMGs shown in Fig. 2. It is easy to confirm that in both ADMGs there exists no valid adjustment set[2] to identify the causal

2. A set Z is said to be a valid adjustment set wrt to T and Y if $p(Y(t)) = \sum_Z p(Y \mid T = t, Z) \times p(Z)$.

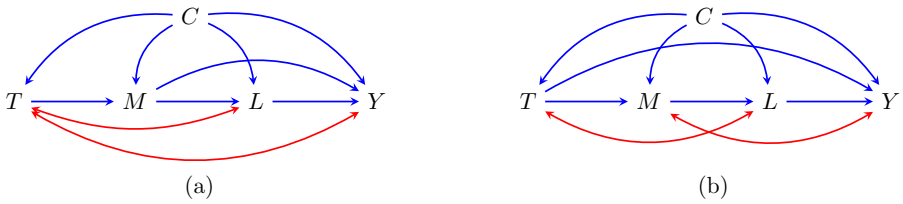


Figure 2: Examples of acyclic directed mixed graphs where T is primal fixable.

effect of T on Y. However, this effect is indeed identified in both graphs via more complicated functionals. The defining characteristic of these ADMGs that permits identification of the target $\psi(t)$ is that the district of T does not intersect with any of its children.

In this section we consider the class of ADMGs where $\text{dis}_{\mathcal{G}}(T) \cap \text{ch}_{\mathcal{G}}(T) = \emptyset$. This criterion encompasses many popular models in the literature, including those that satisfy the back-door and front-door criteria (Pearl, 2009, 1995) as special cases. We name this criterion primal fixability or *p-fixability* for short, due to its generalization of the fixing criterion introduced in the definition of the nested Markov model. In what follows we discuss several identification and estimation methods for the effect of a p-fixable treatment T on outcome Y.

Assume $p(V)$ factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable, and assume, without loss of generality, that Y has no descendants in $\mathcal{G}$.[3] For the remainder of the paper, we assume a fixed valid topological ordering τ where the treatment T appears later than all of its non-descendants i.e., $T \succ_\tau V \setminus \text{de}_{\mathcal{G}}(T)$, and the outcome Y is the final element in the topological ordering. This allows for easier exposition by fixing the definition of pre-treatment covariates as being any variable that appears earlier than T under the ordering τ.

We partition the set of all variables in V into three disjoint sets: (i) all pre-treatment variables, (ii) all post-treatment variables that are in the same district as T, and (iii) all post-treatment variables that are *not* in the district of T. Let D_T denote the district of T.[4] Then V is partitioned as follows: $V = \{\mathbb{C}, \mathbb{L}, \mathbb{M}\}$ where

$$\begin{aligned} \mathbb{C} &= \{C_i \in V \mid C_i \prec T\}, \\ \mathbb{L} &= \{L_i \in V \mid L_i \in D_T, L_i \succeq T\}, \\ \mathbb{M} &= \{M_i \in V \mid M_i \notin \mathbb{C} \cup \mathbb{L}\}. \end{aligned} \tag{13}$$

Primal fixability is known to be a necessary and sufficient condition for the identifiability of the causal effect of T on all other variables $V \setminus T$ (Tian and Pearl, 2002a). In observed data distributions $p(V)$ that district factorize according to an ADMG $\mathcal{G}(V)$ where T is primal fixable, the resulting identifying functional for the target is

$$\psi(t) = \sum_{V \setminus T} Y \times \prod_{M_i \in \mathbb{M}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\Big|_{T=t} \times \sum_{T} \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times p(\mathbb{C}), \tag{14}$$

3. As discussed in Section 5 the efficient IF based estimator is not a function of descendants of Y.
4. The special notation for the district of T as D_T is due to its frequent occurrence in subsequent results.

where $\Big|_{T=t}$ denotes the evaluation of $p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))$ at $T = t$, for all $M_i \in \mathbb{M}$.

We now discuss four different estimators for the above identifying functional that rely on parametric models for a subset of pieces in the observed data distribution while allowing the other pieces to remain unrestricted. The four estimators are: (i) plug-in maximum likelihood, (ii) primal IPW, (iii) dual IPW, and (iv) augmented primal IPW; estimators (ii)–(iv) are novel contributions of the present work. Under regularity conditions (e.g., Robins et al. (1992)), all four estimators are consistent and asymptotically normal under the corresponding assumed model, discussed below. We will show that the primal and dual IPW estimators rely on distinct and variationally independent pieces of a natural parameterization for the observed data likelihood, and that augmented primal IPW is doubly robust under a semiparametric union model thereby allowing for robustness to partial model misspecification.

4.1 Semiparametric Estimation

The first of the four estimators is the plug-in maximum likelihood estimator (MLE) in a semiparametric model where conditional densities for variables in $\mathbb{M} \cup \mathbb{L}$ are assumed parametric forms and the law of $\mathbb{C}$ is unrestricted. We denote this statistical model by $\mathcal{M}_{\mathbb{M}\cup\mathbb{L}} := \{p(V;\eta) = p(\mathbb{C}) \times \prod_{V_i \in V\setminus\mathbb{C}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i);\eta_{v_i}) : p(\mathbb{C}) \text{ is unrestricted and } \eta_{v_i} \in \Gamma_i\}$, with η_{v_i} indexing a finite set of parameters in the parameter space Γ_i.

The parametric factors could in principle be made as flexible as allowed by sample size. By the plug-in principles, the MLE when $p(V) \in \mathcal{M}_{\mathbb{M}\cup\mathbb{L}}$, is obtained via

$$\widehat{\psi(t)}_{\text{mle}} = \mathbb{P}_n\left[\sum_{V\setminus T,\mathbb{C}} Y \times \prod_{M_i\in\mathbb{M}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i);\widehat{\eta}_{m_i})\Big|_{T=t} \times \sum_{T}\prod_{L_i\in\mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i);\widehat{\eta}_{l_i})\right], \quad (15)$$

where $\mathbb{P}_n[.] := 1/n\sum_{j=1}^{n}(.)$ and $\widehat{\eta}_{v_i}$ denotes the MLE of η_{v_i}. We can simplify the above estimator based on how Y is related to T : If $Y \in D_T$, then $Y \in \mathbb{L}$ and $\sum_Y Y \times p(Y \mid \text{mp}_{\mathcal{G}}(Y)) = \mathbb{E}[Y \mid \text{mp}_{\mathcal{G}}(Y)]$ evaluated at the observed value of T; if $Y \notin D_T$, then $Y \in \mathbb{M}$ and $\sum_Y Y \times p(Y \mid T = t, \text{mp}_{\mathcal{G}}(Y)\setminus T) = \mathbb{E}[Y \mid T = t, \text{mp}_{\mathcal{G}}(Y)\setminus T]$. The estimator in (15) is only consistent under the correct specification of the required models for variables in $\mathbb{M}\cup\mathbb{L}$, i.e., when $p(V) \in \mathcal{M}_{\mathbb{M}\cup\mathbb{L}}$.

We can further simplify the estimator in (15), by grouping T and $\mathbb{C}$ together, and evaluating the joint distribution $p(\mathbb{C}, T)$ empirically. We refer to the corresponding statistical model by $\mathcal{M}_{\mathbb{M}\cup\mathbb{L}\setminus T}$, which is a supermodel of $\mathcal{M}_{\mathbb{M}\cup\mathbb{L}}$ where $p(T \mid \mathbb{C})$ is also left unrestricted. The MLE when $p(V) \in \mathcal{M}_{\mathbb{M}\cup\mathbb{L}\setminus T}$ is given by

$$\widehat{\psi(t)}_{\text{mle},2} = \mathbb{P}_n\left[\sum_{V\setminus T,\mathbb{C}} Y \times \prod_{M_i\in\mathbb{M}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i);\widehat{\eta}_{m_i})\Big|_{T=t} \times \prod_{L_i\in\mathbb{L}\setminus T} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i);\widehat{\eta}_{l_i})\right].$$

4.2 Primal and Dual IPW Estimators

In this section we introduce new IPW estimators whose consistency rely on correct specification of only particular subsets of models. Since these estimators provide different perspectives on estimating the same target, we draw inspiration from the optimization literature

(Dantzig et al., 1956; Boyd and Vandenberghe, 2004) in naming them primal and dual IPW. We derived these estimators by rethinking the identifying functional given in Eq. 14. We also generalized the original definition of the fixing operator – introduced in Section 2 – to a primal fixing operator, where the new kernel is now derived using the weights from primal IPW. We defer a description of the primal fixing operator to Appendix F. We will first formalize our results, and then discuss some intuition and examples.

Lemma 4 (Primal IPW formulation)
Given a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable, $\psi(t) = \mathbb{E}[\beta(t)_{primal}]$ where

$$\beta(t)_{primal} \equiv \frac{\mathbb{I}(T=t)}{q_{D_T}(T \mid \text{mb}_{\mathcal{G}}(T))} \times Y = \mathbb{I}(T=t) \times \frac{\sum_{T} \prod_{V_i \in \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))}{\prod_{V_i \in \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))} \times Y. \quad (16)$$

The kernel $q_{D_T}(T \mid \text{mb}_{\mathcal{G}}(T))$ in Lemma 4 may be viewed as a *nested* propensity score derived from the post-intervention distribution $q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))$ where all variables outside of D_T are intervened on and held fixed to some constant value. Recall that the kernel $q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))$ is identified as $\prod_{V_i \in D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))$ as in Eq. 6. Consequently, $q_{D_T}(T \mid \text{mb}_{\mathcal{G}}(T))$ is identified by the definition of conditioning on all elements in D_T outside of T in the kernel $q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))$ as,

$$\begin{aligned} q_{D_T}(T \mid \text{mb}_{\mathcal{G}}(T)) &= q_{D_T}(T \mid D_T \cup \text{pa}_{\mathcal{G}}(D_T) \setminus T) = \frac{q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))}{q_{D_T}(D_T \setminus T \mid \text{pa}_{\mathcal{G}}(D_T))} \\ &= \frac{q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))}{\sum_T q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))} = \frac{\prod_{V_i \in D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))}{\sum_T \prod_{V_i \in D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))}. \end{aligned}$$

The final expression simplifies further by noticing that all vertices appearing prior to T under the topological order τ, do not contain T in their Markov pillows. Consequently, $p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))$ is not a function of T if $V_i \prec T$. Thus, these terms may be pulled out of the summation in the denominator, and cancel with the corresponding term in the numerator. This reduces $V_i \in D_T$ to simply $V_i \in \mathbb{L}$, where $\mathbb{L}$ is defined in display (13), and yields the resulting primal IPW formulation in Eq. 16.

We now introduce the dual formulation of $\psi(t)$ in Eq. 14. Define the *inverse Markov pillow* of T as the set of variables outside the district of T that have T in their Markov pillow. Given the definition of $\mathbb{M}$ in display (13), the inverse Markov pillow is going to be a subset of $\mathbb{M}$. We denote this subset by $\mathbb{M}^*$ and define it formally as follows:

$$\mathbb{M}^* = \{V_i \in \mathbb{M} \mid T \in \text{mp}_{\mathcal{G}}(V_i)\}. \quad (17)$$

Lemma 5 (Dual IPW formulation)
Given a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable, $\psi(t) = \mathbb{E}[\beta(t)_{dual}]$ where

$$\beta(t)_{dual} = \frac{\prod_{V_i \in \mathbb{M}^*} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\Big|_{T=t}}{\prod_{V_i \in \mathbb{M}^*} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))} \times Y. \tag{18}$$

The reason why the products in Eq. 18 are over only a subset of variables in $\mathbb{M}$ is straightforward: if there exists $V_j \in \mathbb{M} \setminus \mathbb{M}^*$ – in other words $T \notin \text{mp}_{\mathcal{G}}(V_j)$ – then $p(V_j \mid \text{mp}_{\mathcal{G}}(V_j))|_{T=t}$ in the numerator cancels out with $p(V_j \mid \text{mp}_{\mathcal{G}}(V_j))$ in the denominator.

The representation of $\psi(t)$ as $\mathbb{E}[\beta(t)_{\text{primal}}]$ and $\mathbb{E}[\beta(t)_{\text{dual}}]$ in Lemmas 4 and 5 immediately yields the corresponding primal and dual IPW estimators. Consider the following two semiparametric models:

(i) $\mathcal{M}_{\mathbb{L}}$: where the conditional densities of $p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))$, $\forall L_i \in \mathbb{L}$, assume parametric forms, and everything else in the observed data distribution is unrestricted.

(ii) $\mathcal{M}_{\mathbb{M}}$: where the conditional densities of $p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))$, $\forall M_i \in \mathbb{M}$, assume parametric forms, and everything else in the observed data distribution is unrestricted.

The primal IPW estimator $\widehat{\psi(t)}_{\text{primal}}$ and the dual IPW estimator $\widehat{\psi(t)}_{\text{dual}}$ are obtained as follows:

$$\widehat{\psi(t)}_{\text{primal}} = \mathbb{P}_n \left[\mathbb{I}(T = t) \times \frac{\sum_T \prod_{V_i \in \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i); \widehat{\eta}_{v_i})}{\prod_{V_i \in \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i); \widehat{\eta}_{v_i})} \times Y \right], \tag{19}$$

$$\widehat{\psi(t)}_{\text{dual}} = \mathbb{P}_n \left[\frac{\prod_{V_i \in \mathbb{M}^*} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i); \widehat{\eta}_{v_i})\Big|_{T=t}}{\prod_{V_i \in \mathbb{M}^*} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i); \widehat{\eta}_{v_i})} \times Y \right]. \tag{20}$$

Theorem 6 (Primal and Dual IPW estimators)
Under standard regularity conditions and positivity assumptions, $\widehat{\psi(t)}_{primal}$ and $\widehat{\psi(t)}_{dual}$ are consistent and asymptotically normal under models $\mathcal{M}_{\mathbb{L}}$ and $\mathcal{M}_{\mathbb{M}}$, respectively.

Note that when $Y \in \mathbb{L}$, we can use a weaker set of assumptions for $\widehat{\psi(t)}_{\text{primal}}$ by replacing $p(Y \mid \text{mp}_{\mathcal{G}}(Y); \widehat{\eta}_y) \times Y$ in the numerator with $\mathbb{E}[Y \mid \text{mp}_{\mathcal{G}}(Y); \widehat{\eta}_y]$ and deleting the factor $p(Y \mid \text{mp}_{\mathcal{G}}(Y); \widehat{\eta}_y)$ in the denominator. With this replacement, the consistency of the estimator requires correct specification of the conditional mean of Y rather than its entire distribution. Likewise, when $Y \in \mathbb{M}^*$, we can use weaker assumptions for $\widehat{\psi(t)}_{\text{dual}}$ with a similar move; see Section 4.2.1 for examples. These moves are general and can be applied anytime $Y \in \mathbb{L}$ or $Y \in \mathbb{M}^*$.

The sets of nuisance models in the primal and dual IPW estimators form variationally independent components of natural parameterizations of the observed data distribution $p(V)$, as formalized in the following theorem.

Theorem 7 (Variational independence of primal IPW and dual IPW) *Given a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable, the IPW estimators ψ_{primal} and ψ_{dual} proposed in Lemmas 4 and 5 respectively, use variationally independent components of the observed distribution $p(V)$, under any parameterization of $p(V)$ that is decomposable with respect to districts in $\mathcal{G}(V)$.*

Primal IPW can be viewed as a generalization of the truncated factorization in DAGs to *truncated kernel factorization* in ADMGs. The g-computation algorithm for a DAG model involves truncation of the DAG factorization, namely dropping a simple conditional factor of the treatment given its parents, i.e., $p(V(t)) = \{p(V)/p(T = t \mid \text{pa}_{\mathcal{G}}(T))\}|_{T=t}$. Similarly, the primal formulation can be viewed as truncation of the district factorization in Eq. 14, where the nested conditional factor for the treatment given its Markov blanket, which is a part of one of the district terms in Eq. 14, is dropped from the observed joint distribution, i.e., $p(V(t)) = \{p(V)/q_{D_T}(T \mid \text{mb}_{\mathcal{G}}(T))\}|_{T=t}$. The intuition for the dual IPW can be gained by viewing it as a probabilistic formalization of the node splitting operation in single world intervention graphs (SWIGs) described in Richardson and Robins (2013). To provide more concrete intuition on the primal and dual IPW estimators, we discuss their application to the ADMGs shown in Fig. 2.

4.2.1 Examples: primal and dual IPW estimators

Consider the ADMG in Fig. 2(a). T is primal fixable as there is no bidirected path from T to any of its children, namely M. The inverse Markov pillow of T in Fig. 2(a) is just M. Per Lemmas 4 and 5, the primal and dual IPW formulations for the identifiable functional of the target parameter $\psi(t)$ in Fig. 2(a) are given by,

$$\text{(Fig. 2a)} \qquad \psi(t)_{\text{primal}} = \mathbb{E}\left[\mathbb{I}(T = t) \times \frac{\sum_T \; p(T \mid C) \times p(L \mid T, M, C) \times p(Y \mid T, M, L, C)}{p(T \mid C) \times p(L \mid T, M, C) \times p(Y \mid T, M, L, C)} \times Y \right], \text{ and}$$
$$\psi(t)_{\text{dual}} = \mathbb{E}\left[\frac{p(M \mid T = t, C)}{p(M \mid T, C)} \times Y \right].$$

To estimate $\psi(t)$ we proceed as follows. In case of primal IPW, we fit conditional densities $p(T \mid C), p(L \mid T, M, C)$, and $p(Y \mid T, M, L, C)$, either parametrically (using generalized linear models for instance), or via more flexible models (like generalized additive models or nonparametric kernel regression methods) as long as sample size allows. The target parameter is then obtained by empirically evaluating the outer expectation using the fitted models. We can also avoid modeling the conditional density of Y, as the outcome regression $\mathbb{E}[Y \mid T, M, L, C]$ suffices to estimate $\psi(t)$, i.e., ψ_{primal} can be expressed equivalently as

$$\mathbb{E}\left[\mathbb{I}(T = t) \times \frac{\sum_T \; p(T \mid C) \times p(L \mid T, M, C) \times \mathbb{E}[Y \mid T, M, L, C]}{p(T \mid C) \times p(L \mid T, M, C)} \right].$$

A simple procedure to estimate the dual IPW involves modeling the conditional density $p(M \mid T, C)$. However, a more sophisticated procedure may take advantage of modeling the density ratio directly as suggested by Sugiyama et al. (2010).

We now turn our attention to the ADMG in Fig. 2(b). The inverse Markov pillow of T in Fig. 2(b) is $\{M, Y\}$. The corresponding primal and dual IPW formulations are given by,

(Fig. 2b)
$$\psi_{\text{primal}} = \mathbb{E}\left[\mathbb{I}(T=t) \times \frac{\sum_T p(T \mid C) \times p(L \mid T, M, C)}{p(T \mid C) \times p(L \mid T, M, C)} \times Y\right], \text{ and}$$
$$\psi_{\text{dual}} = \mathbb{E}\left[\frac{p(M \mid T=t, C)}{p(M \mid T, C)} \times \frac{p(Y \mid T=t, M, L, C)}{p(Y \mid T, M, L, C)} \times Y\right].$$

Similar strategies as n the previous case can be used to estimate $\psi(t)$. The conditional density of Y in ψ_{dual} can be replaced vby the outcome regression $\mathbb{E}[Y \mid T=t, M, L, C]$, i.e., ψ_{dual} can be expressed equivalently as

$$\mathbb{E}\Big[\frac{p(M \mid T=t, C)}{p(M \mid T, C)} \times \mathbb{E}[Y \mid T=t, M, L, C]\Big].$$

4.3 Augmented Primal IPW Estimators

In the previous section we have shown the existence of two estimators for the target $\psi(t)$ that use variationally independent portions of the likelihood when T is p-fixable. The question naturally arises if it is possible to combine these estimators to yield a single estimator that exhibits double robustness in the sets of models used in each one. In the following theorem, we derive the efficient influence function for $\psi(t)$ in the nonparametric model, $\mathcal{M}_{\text{np}}$, with no restriction on the observed data distribution. We then prove that the estimator obtained via this influence function is doubly robust. This influence function can be viewed as augmenting the primal IPW with pieces from the dual IPW. For readability, we use $\prod_{L_i \prec M_i}$ as shorthand for $\prod_{L_i \in \mathbb{L} | L_i \prec M_i}$. The sets $\mathbb{C}, \mathbb{L}, \mathbb{M}$ are defined in display (13).

Theorem 8 (The efficient influence function of $\psi(t)$ in $\mathcal{M}_{\text{np}}$)
Given a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable, the efficient influence function for the target parameter $\psi(t)$ in the nonparametric model $\mathcal{M}_{np}$ is as follows.

$$\begin{aligned}
U_{\psi_t} = \sum_{M_i \in \mathbb{M}} \Bigg\{ & \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_i} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \Bigg(\sum_{T \cup \{\succ M_i\}} Y \times \prod_{\substack{V_i \in \mathbb{L} \cup \\ \{\succ M_i\}}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\Big|_{T=t \text{ if } V_i \in \mathbb{M}} \\
& - \sum_{T \cup \{\succeq M_i\}} Y \times \prod_{\substack{V_i \in \mathbb{L} \cup \\ \{\succeq M_i\}}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\Big|_{T=t \text{ if } V_i \in \mathbb{M}} \Bigg) \Bigg\} \\
+ \sum_{L_i \in \mathbb{L} \setminus T} \Bigg\{ & \frac{\prod_{M_i \prec L_i} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\Big|_{T=t}}{\prod_{M_i \prec L_i} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times \Bigg(\sum_{\{\succ L_i\}} Y \times \prod_{V_i \succ L_i} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\Big|_{T=t \text{ if } V_i \in \mathbb{M}} \\
& - \sum_{\{\succeq L_i\}} Y \times \prod_{V_i \succeq L_i} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\Big|_{T=t \text{ if } V_i \in \mathbb{M}} \Bigg) \Bigg\} \\
+ \sum_{V \setminus \{T, \mathbb{C}\}} & Y \times \prod_{M_i \in \mathbb{M}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\Big|_{T=t} \times \prod_{L_i \in \mathbb{L} \setminus T} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) - \psi(t), \qquad (21)
\end{aligned}$$

where the sets $\mathbb{C}, \mathbb{L}, \mathbb{M}$ are defined in display (13). The asymptotic efficiency bound is given by the variance of U_{ψ_t}.

The efficient influence function contains three terms. The first may be viewed as a centered version of the primal IPW estimator (16), the second as a centered version of the dual IPW estimator (18), and the third as a centered version of the plug-in estimator (14).

The influence function U_{ψ_t} in Theorem 8 depends on unknown conditional densities (a.k.a. nuisance parameters). Let $\widehat{U}_{\psi_t}$ denote the influence function where the unknown nuisance parameters are replaced with their corresponding estimators. Thus, the estimating equation $\mathbb{P}_n[\widehat{U}_{\psi_t}] = 0$ yields an estimator for $\psi(t)$ that we call *augmented primal IPW* (APIPW). In the following theorem, we show that this estimator exhibits a double robustness behavior with respect to models involving variables in $\mathbb{M}$ and $\mathbb{L}$.

Theorem 9 (Double robustness of Augmented Primal IPW)
Under standard regularity conditions and positivity assumptions, the estimator obtained by solving the estimating equation $\mathbb{P}_n[\widehat{U}_{\psi_t}] = 0$, where U_{ψ_t} is given in Theorem 8, is consistent and asymptotically normal if all models in either $\{p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)),\ \forall M_i \in \mathbb{M}\}$ or $\{p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)),\ \forall L_i \in \mathbb{L}\}$ are correctly specified. Further, the estimator is locally semiparametric efficient for the functional $\psi(t)$ under the union model $\mathcal{M}_{\mathbb{M}} \cup \mathcal{M}_{\mathbb{L}}$, at the intersection submodel $\mathcal{M}_{\mathbb{M}} \cap \mathcal{M}_{\mathbb{L}}$.

According to Theorem 9, the APIPW estimator is a *doubly robust* estimator. This allows us to perform consistent inference for the target parameter $\psi(t)$ even in settings where a large part of the model likelihood is arbitrarily misspecified, provided that the appropriate conditional models for variables in either $\mathbb{M}$ or $\mathbb{L}$ are specified correctly. The double robustness of the APIPW estimator stems from the fact that its bias has a product form which allows parametric $(\sqrt{n})$ convergence rates for $\psi(t)$ to be obtained even if flexible machine learning models with slower than parametric convergence rates (but faster than $n^{-1/4}$) are used to fit the nuisance models, that is the conditional factors involving variables in $\mathbb{M}$ and $\mathbb{L}$; see Chernozhukov et al. (2018) for more details and Robins et al. (2008); Rotnitzky et al. (2020) for further discussions.

4.3.1 Cancellation of Terms in the IF

Given a post treatment variable V_i and its conditional density $p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))$ in the identified functional for $\psi(t)$ in Eq. 14, there is a corresponding term in the influence function U_{ψ_t} in Theorem 8 of the form

$$f_1(\prec V_i) \times \Big(f_2(\preceq V_i) - \sum_{V_i} f_2(\preceq V_i) \times p\big(V_i \mid \text{mp}_{\mathcal{G}}(V_i)\big) \Big), \tag{22}$$

where $f_1(\prec V_i)$ denotes a function of variables that precedes V_i in the topological order. Similarly, $f_2(\preceq V_i)$ is a function of $\{\prec V_i\}$ and V_i itself. Sometimes, these terms in the influence function U_{ψ_t} may cancel each other out. For instance, assume there are two consecutive variables $V_i, V_{i+1} \in \mathbb{L}$ (or $\in \mathbb{M}$) such that $\text{mp}_{\mathcal{G}}(V_{i+1}) \setminus V_i \subseteq \text{mp}_{\mathcal{G}}(V_i)$. The corresponding terms in the influence function share some common terms: First, the two share the same weight terms, i.e., $f_1(\prec V_{i+1}) = f_1(\prec V_i)$, and second $f_2(\preceq V_i) = \sum_{V_{i+1}} f_2(\preceq V_{i+1}) \times p\big(V_{i+1} \mid \text{mp}_{\mathcal{G}}(V_{i+1})\big)$. Therefore, through simple algebra, we note that V_i and V_{i+1} can be viewed as contributing a single term to the influence function of the form shown in Eq. 22, and that is

$$f_1(\prec V_{i+1}) \times \Big(f_2(\preceq V_{i+1}) - \sum_{V_i, V_{i+1}} f_2(\preceq V_{i+1}) \times p\big(V_{i+1}, V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)\big) \Big).$$

This cancellation occurs regardless of whether $V_i \in \mathbb{L}$ or $V_i \in \mathbb{M}$. However it is important that both V_i and V_{i+1} be in the same set (since they need a common f_1 term to be factored out.) Such cancellations may be applied recursively to consecutive variables in $\mathbb{L}$ or $\mathbb{M}$.

Another possible cancellation of terms may occur in the weights that correspond to "dual weights" in Eq. 21. The factors in the numerator and the denominator are exactly the same except for the fact that the numerator is evaluated at $T = t$. However, if there exists $M_i \in \mathbb{M}$ such that T is not in its Markov pillow, i.e., $M_i \perp\!\!\!\perp T \mid \mathrm{mp}_{\mathcal{G}}(M_i)$, then $p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} / p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) = 1$. Note that such cancellations only involves variables in $\mathbb{M}$. In fact, the set of conditional densities that stay in these weight terms correspond to the variables in the inverse Markov pillow of T, previously denoted by $\mathbb{M}^*$. Therefore, we have the following general simplification to the influence function U_{ψ_t} in Theorem 8,

$$\frac{\prod\limits_{M_i \prec L_i} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))\Big|_{T=t}}{\prod\limits_{M_i \prec L_i} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} = \frac{\prod\limits_{M_i \in\, \mathbb{M}^* \,\cap\, \{\prec L_i\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))\Big|_{T=t}}{\prod\limits_{M_i \in\, \mathbb{M}^* \,\cap\, \{\prec L_i\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))}.$$

More intuition on the nonparametric IF is provided in Appendix E. An implication of the two aforementioned forms of cancellation is that the robustness statement in Theorem 9 is somewhat conservative. In other words, it may not be necessary to model all the conditional terms mentioned in the doubly robust statement of Theorem 9. It is sometimes possible to prune vertices from the ADMG and still achieve a doubly robust estimator that requires fitting less models as demonstrated via an example in the next subsection.

4.3.2 Reformulation of the IF

An alternative representation of the influence function in Theorem 8 can be expressed solely in terms of the primal and dual IPW statements in Lemmas 4 and 5. This formulation serves as a helpful analytic tool for deriving efficient IFs in Section 5, and can have practical implications in terms of practical usage.

Lemma 10 (Reformulation of the IF for augmented primal IPW)
Under the same conditions stated in Theorem 8, the efficient influence function for the target parameter $\psi(t)$ in the nonparametric model $\mathcal{M}_{np}$ can be re-expressed as follows.

$$\begin{aligned} U_{\psi_t} = &\sum_{M_i \in \mathbb{M}} \mathbb{E}[\beta_{primal} \mid \{\preceq M_i\}] - \mathbb{E}[\beta_{primal} \mid \{\prec M_i\}] \\ &+ \sum_{L_i \in \mathbb{L}} \mathbb{E}[\beta_{dual} \mid \{\preceq L_i\}] - \mathbb{E}[\beta_{dual} \mid \{\prec L_i\}] \\ &+ \mathbb{E}[\beta_{primal/dual} \mid \mathbb{C}] - \psi(t), \end{aligned}$$

where β_{primal} and β_{dual} are defined via Lemmas 4 and 5 respectively, and $\beta_{primal/dual}$ means that we may use either β_{primal} or β_{dual}.

According to the above lemma, the portion of the IF that relates to elements in $\mathbb{C}$, may be recovered using either β_{primal} or β_{dual}. That is, $\mathbb{E}[\beta_{\text{primal}} \mid \mathbb{C}] = \mathbb{E}[\beta_{\text{dual}} \mid \mathbb{C}]$ and $\mathbb{E}[\beta_{\text{primal}}] = \mathbb{E}[\beta_{\text{dual}}] = \psi(t)$. Reformulation of the IF offers the advantage of restricting the modeling of conditional densities to only those involved in β_{primal} and β_{dual}. The analyst may then rely on flexible regression methods in order to model each $\mathbb{E}[\cdot \mid \cdot]$ above. The downside of such a formulation is that we may no longer be able to take advantage of the double robustness properties of the APIPW estimators, as stated in Theorem 9. However, such a reformulation is quite useful for deriving the form of efficient IFs by performing projections onto the tangent space of the model; we provide further comments on estimation in Section 5 after deriving efficient IFs in mb-shielded ADMGs. From a practical stand-point, it is easy to translate back and forth between the reformulation and the original IF using Theorem 8 and Lemma 10 depending on the needs of the data analyst.

4.3.3 Examples: Augmented Primal IPW

We now revisit the ADMGs in Fig. 2 and derive the corresponding efficient influence functions in the nonparametric model $\mathcal{M}_{\text{np}}$. Consider the ADMG in Fig. 2(a). The sets in display (13) are as follows, $\mathbb{C} = \{C\}, \mathbb{L} = \{T, L, Y\}$, and $\mathbb{M} = \{M\}$. Applying Theorem 8 to this graph, yields the influence function:

$$
\begin{aligned}
\text{(Fig. 2a)} \qquad U_{\psi_t} = \frac{\mathbb{I}(T=t)}{p(T \mid C)} \times \Big(& \sum_{T,L} p(T \mid C) \times p(L \mid T, M, C) \times \mathbb{E}[Y \mid T, M, L, C] \\
& - \sum_{T,L,M} p(M \mid T = t, C) \times p(T \mid C) \times p(L \mid T, M, C) \times \mathbb{E}[Y \mid T, M, L, C] \Big) \\
& + \frac{p(M \mid T = t, C)}{p(M \mid T, C)} \times \Big(Y - \mathbb{E}[Y \mid T, M, L, C] \Big) \\
& + \frac{p(M \mid T = t, C)}{p(M \mid T, C)} \times \Big(\mathbb{E}[Y \mid T, M, L, C] \; - \sum_{L} p(L \mid T, M, C) \times \mathbb{E}[Y \mid T, M, L, C] \Big) \\
& + \sum_{M,L} p(M \mid T = t, C) \times p(L \mid T, M, C) \times \mathbb{E}[Y \mid T, M, L, C] - \psi(t).
\end{aligned}
$$

Note that in the above influence function, the term $\frac{p(M|T=t,C)}{p(M|T,C)} \times \mathbb{E}[Y \mid T, M, L, C]$ appears twice with opposite signs. This is an example of the kind of cancellation mentioned in the previous section, where Y and L are consecutive elements in the set $\mathbb{L}$ that share essentially the same Markov pillow, i.e., $\text{mp}_{\mathcal{G}}(Y) \setminus L = \text{mp}_{\mathcal{G}}(L)$. In fact, this observation allows us to simplify the influence function even further by deriving the influence function in the ADMG $\mathcal{G}(V \setminus L)$ where L is treated as latent; projecting out L in this example, corresponds to removing all the edges into and out of L. This ADMG is simply the front-door graph with baseline confounding. Given Theorem 8, the IF is as follows.

$$
\begin{aligned}
\text{(Fig. 2a)} \quad U_{\psi_t} = \frac{\mathbb{I}(T=t)}{p(T \mid C)} \times \Big(& \sum_{T} p(T \mid C) \times \mathbb{E}[Y \mid T, M, C] \; - \sum_{T,M} p(M \mid T = t, C) \times p(T \mid C) \times \mathbb{E}[Y \mid T, M, C] \Big) \\
& + \frac{p(M \mid T = t, C)}{p(M \mid T, C)} \times \Big(Y - \mathbb{E}[Y \mid T, M, C] \Big) \; + \sum_{M} p(M \mid T = t, C) \times \mathbb{E}[Y \mid T, M, C] - \psi(t).
\end{aligned}
$$

Now consider the ADMG in Fig. 2(b) where no simplification of the IF is possible. The sets in display (13) are $\mathbb{C} = \{C\}, \mathbb{L} = \{T, L\}$, and $\mathbb{M} = \{M, Y\}$. The influence function per Theorem 8 is,

$$
\begin{aligned}
\text{(Fig. 2b)} \qquad U_{\psi_t} = \frac{\mathbb{I}(T=t)}{p(T \mid C) \times p(L \mid T, M, C)} \times \Big(& \sum_{T} p(T \mid C) \times p(L \mid T, M, C) \times Y \\
& - \sum_{T} p(T \mid C) \times p(L \mid T, M, C) \times \mathbb{E}[Y \mid T = t, M, L, C] \Big) \\
+ \frac{\mathbb{I}(T=t)}{p(T \mid C)} \times \Big(& \sum_{T,L} p(T \mid C) \times p(L \mid T, M, C) \times \mathbb{E}[Y \mid T = t, M, L, C] \\
& - \sum_{T,M,L} p(T \mid C) \times p(M \mid T = t, C) \times p(L \mid T, M, C) \times \mathbb{E}[Y \mid T = t, M, L, C] \Big) \\
+ \frac{p(M \mid T = t, C)}{p(M \mid T, C)} \times \Big(& \mathbb{E}[Y \mid T = t, M, L, C] \; - \sum_{L} p(L \mid T, M, C) \times \mathbb{E}[Y \mid T = t, M, L, C] \Big) \\
+ \sum_{M,L} & p(M \mid T = t, C) \times p(L \mid T, M, C) \times \mathbb{E}[Y \mid T = t, M, L, C] - \psi(t). \qquad (23)
\end{aligned}
$$

We briefly describe estimation strategies for estimators resulting from the Theorem 8 using the influence function in Eq. 23 as an example. An estimator for the target $\psi(t)$ is obtained by solving the estimating equation $\mathbb{P}_n[U_{\psi_t}] = 0$. In the resulting estimator, conditional densities for $p(T \mid C), p(M \mid T, C), p(L \mid T, M, C)$ and the outcome regression $\mathbb{E}[Y \mid T, M, L, C]$ can be fit either parametrically or using flexible machine models (as long as the rates of convergence are fast enough to achieve $\sqrt{n}$-consistency). The outer expectation is then evaluated empirically using the fitted models in order to yield the target parameter. Per Theorem 9, the estimator for $\psi(t)$ is consistent as long as one of the sets $\{p(T \mid C), p(L \mid T, M, C)\}$ or $\{p(M \mid T, C), \mathbb{E}[Y \mid T, M, L, C]\}$ is correctly specified while allowing for arbitrary misspecification of the other.

Another estimation strategy that is computationally simpler stems from the usage of Theorem 10 to the ADMG in Fig. 2(b). With the simplification that $\mathbb{E}[\beta_{\text{primal}} \mid Y, T, M, L, C] = \beta_{\text{primal}}$, the target can be written as follows.

$$
\begin{aligned}
\text{(Fig. 2b)} \quad \psi(t)_{\text{reform}} = \; \mathbb{E}\big[& \beta_{\text{primal}} - \mathbb{E}[\beta_{\text{primal}} \mid T, M, L, C] \\
& + \mathbb{E}[\beta_{\text{primal}} \mid M, T, C] - \mathbb{E}[\beta_{\text{primal}} \mid T, C] \\
& + \mathbb{E}[\beta_{\text{dual}} \mid L, T, M, C] - \mathbb{E}[\beta_{\text{dual}} \mid T, M, C] \\
& + \mathbb{E}[\beta_{\text{dual}} \mid T, C] \,\big]. \qquad (24)
\end{aligned}
$$

The above can be estimated from finite samples by first obtaining estimates for β_{primal} and β_{dual} for each row in our data and then fitting flexible regressions for each $\mathbb{E}[\cdot \mid \cdot]$ shown in Eq. 24 using these estimates as pseudo outcomes. The outer expectation is then evaluated empirically using these fitted models, yielding an estimate for the target parameter $\psi(t)$.

The two estimation strategies described above come with trade-offs. The former approach requires modeling conditional densities and computing sums, but preserves the double robustness property and does not face issues of model compatibility. The latter approach trades model compatibility and double robustness for computational efficiency.

4.4 Special Case of Simplification When the Treatment Is Fixable

Consider the class of ADMGs where in addition to being primal fixable, the treatment T has no bidirected path to any of its descendants, i.e., $\text{dis}_{\mathcal{G}}(T) \cap \text{de}_{\mathcal{G}}(T) = \{T\}$. This coincides with the original criterion of fixing used in the definition of the nested Markov model in Section 2. We now show that when this condition holds that the identification functional in Eq. 14 simplifies to covariate adjustment using the Markov pillow of the treatment, and the corresponding influence function then simplifies to standard augmented IPW. Recall that we are assuming a fixed valid topological ordering τ where the treatment T occurs after all its non-descendants. When the treatment is fixable, it is easy to show that this implies the Markov pillow of T and Markov blanket of T are the same, $\text{mp}_{\mathcal{G}}(T) = \text{mb}_{\mathcal{G}}(T)$. The following result follows by definition of fixing, m-separation, and the backdoor adjustment criterion (Pearl, 1995).

Lemma 11 (Identifying functional when T is fixable)
Given a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ in which T is fixable, $\psi(t)$ is identified as $\psi(t) = \mathbb{E}[\mathbb{E}[Y \mid T = t, \text{mp}_{\mathcal{G}}(T)]]$.

The identifiability of the target in this manner, immediately yields that the efficient influence function of Theorem 8 simplifies to the one corresponding to the AIPW estimator except the conditioning set is now extended to include members of the district of T and parents of this district. While the above result is not an exhaustive criterion for when such simplifications occur, it provides a simple link between the fixing operator used to define nested Markov models and the validity of covariate adjustment. For general criteria regarding covariate adjustment in the presence of hidden variables we refer the reader to Shpitser et al. (2010); Perković et al. (2015).

5. Semiparametric Efficiency Bounds

In Section 3, we provided Algorithm 1 as a means of checking whether the model implied by an ADMG $\mathcal{G}(V)$ is nonparametrically saturated. In an NPS model with a p-fixable treatment, the augmented primal IPW estimator is not only doubly robust but also the most efficient estimator. On the other hand, constraints in a semiparametric model shrink the tangent space of the model. Hence, we no longer have a *unique* influence function. (the class of all influence functions is $\{U_\psi + \Lambda^\perp\}$). In this section, we discuss efficiency results for the class of mb-shielded ADMGs that was proposed in Theorem 2. The general form of the efficient IF in an arbitrary mb-shielded ADMG where T is p-fixable is provided in the following theorem.

Theorem 12 (Efficient augmented primal IPW in mb-shielded ADMGs)
Given a distribution $p(V)$ that district factorizes with respect to an mb-shielded ADMG $\mathcal{G}(V)$ where T is primal fixable, the efficient influence function for the target parameter $\psi(t)$ is given as follows,

$$U_{\psi_t}^{eff} = \sum_{M_i \in \mathbb{M}} \mathbb{E}[\beta_{primal} \mid M_i, \text{mp}_{\mathcal{G}}(M_i)] - \mathbb{E}[\beta_{primal} \mid \text{mp}_{\mathcal{G}}(M_i)]$$

$$+ \sum_{L_i \in \mathbb{L}} \mathbb{E}[\beta_{dual} \mid L_i, \text{mp}_{\mathcal{G}}(L_i)] - \mathbb{E}[\beta_{dual} \mid \text{mp}_{\mathcal{G}}(L_i)]$$
$$+ \sum_{C_i \in \mathbb{C}} \mathbb{E}[\beta_{primal/dual} \mid C_i, \text{mp}_{\mathcal{G}}(C_i)] - \mathbb{E}[\beta_{primal/dual} \mid \text{mp}_{\mathcal{G}}(C_i)] \quad (25)$$

where $\mathbb{C}, \mathbb{L}, \mathbb{M}$ are defined in display (13), and β_{primal} and β_{dual} are obtained as in Lemmas 4 and 5 respectively. $\beta_{primal/dual}$ means that we can either use β_{primal} or β_{dual}.

The primal and dual IPWs comprise the fundamental elements of the efficient influence function in the setting where T is primal fixable. Simplified symbolic representations of the efficient IF in terms of the conditional densities that appear in the topological factorization can be obtained by plugging in the expression from Theorem 8 into computer algebra systems, such as Tikka and Karvanen (2017) and Maxima (2020). Further details on the use of these tools is outside of the scope of the paper, and we will work with the above representation for brevity.

A simple estimation procedure based on Eq. 25 proceeds as follows. First, fit nuisance models for β_{primal} and β_{dual} and obtain estimates of these for each row in the data. Next, fit flexible regressions for each $\mathbb{E}[\cdot \mid \cdot]$ appearing in Eq. 25 using the estimates of β_{primal} and β_{dual} as pseudo outcomes. The estimating equation $\mathbb{P}_n[\widehat{U}^{\text{eff}}_{\psi_t}] = 0$ yields an estimator for the target $\psi(t)$.

In Section 4.4, we discussed when treatment is not just primal fixable but fixable, the identifying functional in Eq. 14 simplified to the adjustment functional and the augmented primal IPW estimator simplified to augmented IPW. Similarly, the efficient IF in Eq. 25 simplies when T is fixable. In the following lemma, we show that the efficient IF in mb-shielded ADMGs where T is fixable is a special case of the efficient IF in Theorem 12, and uses only terms that involve β_{primal}. This result can also be directly obtained from Rotnitzky and Smucler (2020) (since constraints in mb-shielded ADMGs are ordinary conditional independencies.) The conditional independences below rely on a slight abuse of notation where $A \perp\!\!\!\perp B \mid C$ when $B \cap C \neq \emptyset$ is taken to mean $A \perp\!\!\!\perp B \setminus \{B \cap C\} \mid C$.

Lemma 13 (Efficient augmented IPW in mb-shielded ADMGs)
Given a distribution $p(V)$ that district factorizes with respect to an mb-shielded ADMG $\mathcal{G}(V)$ where T is fixable, the efficient influence function for the target parameter $\psi(t)$ is given as follows,

$$U^{eff}_{\psi_t} = \sum_{V_i \in V^*} \mathbb{E}\Big[\frac{\mathbb{I}(T=t)}{p(T \mid \text{mp}_{\mathcal{G}}(T))} \times Y \;\Big|\; V_i, \text{mp}_{\mathcal{G}}(V_i)\Big] - \mathbb{E}\Big[\frac{\mathbb{I}(T=t)}{p(T \mid \text{mp}_{\mathcal{G}}(T))} \times Y \;\Big|\; \text{mp}_{\mathcal{G}}(V_i)\Big],$$

where $V^* = V \setminus (T \cup Z \cup D)$ and

$$Z = \{Z_i \in V \mid Z_i \perp\!\!\!\perp Y \mid \text{mp}_{\mathcal{G}}(Z_i) \text{ in } \mathcal{G}_{V \setminus T} \text{ and } Z_i \not\perp\!\!\!\perp T \mid \text{mp}_{\mathcal{G}}(Z_i)\},$$
$$D = \{D_i \in V \mid D_i \perp\!\!\!\perp T, \text{mp}_{\mathcal{G}}(T), Y \mid \text{mp}_{\mathcal{G}}(D_i)\}.$$

Some interesting facts follow from the form of the efficient influence function shown in Lemma 13. First, the efficient influence function can be obtained by simply projecting the

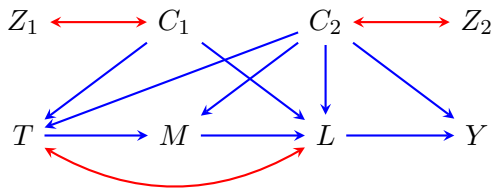


Figure 3: An mb-shielded ADMG that is not NPS and where T is primal fixable.

(primal) IPW portion of the AIPW influence function. Second, the set $V \setminus V^*$ enumerates several vertices that do not affect the efficiency of estimating the target parameter $\psi(t)$. These include vertices Z_i that meet the criteria for a *conditional instrumental variable* (conditional on their Markov pillow) as defined in Pearl (2009); van der Zander et al. (2015). Further, no efficiency is lost by disregarding other vertices D_i that include descendants of Y, and irrelevant non-descendants of Y, as given by definitions of the set D in Lemma 13. For further, extensive discussion of connections between efficient IFs and graphical features in fully observed DAGs see Rotnitzky and Smucler (2020).

5.1. Example: Efficient APIPW

Consider the ADMG shown in Fig. 3. The conditional independencies implied by the graphs are: $C_1 \perp\!\!\!\perp C_2$ and $M \perp\!\!\!\perp C_1, Z_1, Z_2 \mid T, C_2$. As this model is no longer NPS, the IF obtained via Theorem 8 is not the most efficient. However, it is easy to see that this ADMG is mb-shielded and therefore the efficient IF is given by Theorem 12. Fix a valid topological order $(C_1, C_2, Z_1, Z_2, T, M, L, Y)$. We have:

$$\text{(Fig. 3)} \qquad \beta_{\text{primal}} = \mathbb{I}(T = t) \times \frac{\sum_T p(T \mid C_1, C_2) \times p(L \mid T, M, C_1, C_2)}{p(T \mid C_1, C_2) \times p(L \mid T, M, C_1, C_2)} \times Y,$$
$$\beta_{\text{dual}} = \frac{p(M \mid T = t, C_2)}{p(M \mid T, C_2)} \times Y. \tag{26}$$

Define the sets $\mathbb{M} = \{M, Y\}$, $\mathbb{L} = \{T, L\}$, and $\mathbb{C} = \{C_1, C_2\}$. Note that we have dropped terms involving the vertices Z_1 and Z_2 as it is easy to check that $\mathbb{E}[\beta_{\text{dual}} \mid Z_i, \text{mp}_{\mathcal{G}}(Z_i)] = \mathbb{E}[\beta_{\text{dual}} \mid \text{mp}_{\mathcal{G}}(Z_i)]$, resulting in a cancellation of these terms. By applying Theorem 12 to Fig. 3, we get the following form of the efficient APIPW.

$$\begin{aligned}\text{(Fig. 3)} \qquad \psi(t)_{\text{eff-apipw}} = \mathbb{E}\big[&\mathbb{E}[\beta_{\text{primal}} \mid Y, L, C_2] - \mathbb{E}[\beta_{\text{primal}} \mid L, C_2] \\ &+ \mathbb{E}[\beta_{\text{primal}} \mid M, T, C_2] - \mathbb{E}[\beta_{\text{primal}} \mid T, C_2] \\ &+ \mathbb{E}[\beta_{\text{dual}} \mid L, M, T, C_1, C_2] - \mathbb{E}[\beta_{\text{dual}} \mid M, T, C_1, C_2] \\ &+ \mathbb{E}[\beta_{\text{dual}} \mid T, C_1, C_2] - \mathbb{E}[\beta_{\text{dual}} \mid C_1, C_2] \\ &+ \mathbb{E}[\beta_{\text{dual}} \mid C_2] + \mathbb{E}[\beta_{\text{dual}} \mid C_1] - \mathbb{E}[\beta_{\text{dual}}]\big]\end{aligned} \tag{27}$$

The estimation strategy for the above functional is very similar to the one used for Eq. 24 and elaborated upon earlier in the section.

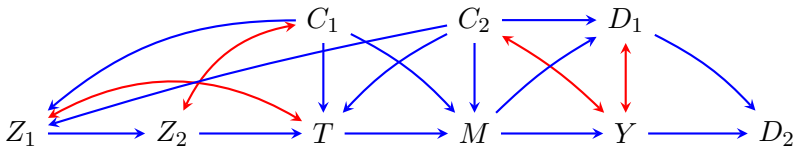


Figure 4: An mb-shielded ADMG that is not NPS and where T is fixable.

5.2 Example: Efficient AIPW

As a concrete example of simplification of the efficient IF under fixability consider the mb-shielded ADMG in Fig. 4. Fix the topological order $\tau = \{C_1, C_2, Z_1, Z_2, T, M, Y, D_1, D_2\}$. One can check that the vertices labeled Z_1 and Z_2 meet the criteria for conditional instruments and the vertices D_1 and D_2 meet the criteria of $D_i \perp\!\!\!\perp T, \text{mp}_{\mathcal{G}}(T), Y \mid \text{mp}_{\mathcal{G}}(D_i)$ and thus do not appear in the terms of the efficient influence function given in Lemma 13. When T is fixable, it is always true that $\beta_{\text{primal}} = \{\mathbb{I}(T = t)/p(T \mid \text{mp}_{\mathcal{G}}(T))\} \times Y$. Consequently, by applying Lemma 13, we obtain the following form of the efficient AIPW for $\psi(t)$.

$$\begin{aligned}\text{(Fig. 4)} \qquad \psi(t)_{\text{eff-aipw}} = \mathbb{E}\Big[& \mathbb{E}[\beta_{\text{primal}} \mid Y, M, C_2] - \mathbb{E}[\beta_{\text{primal}} \mid M, C_2] \\ & + \mathbb{E}[\beta_{\text{primal}} \mid M, T, C_1, C_2] - \mathbb{E}[\beta_{\text{primal}} \mid T, C_1, C_2] \\ & + \mathbb{E}[\beta_{\text{primal}} \mid C_2] + \mathbb{E}[\beta_{\text{primal}} \mid C_1] - \mathbb{E}[\beta_{\text{primal}}] \Big], \end{aligned} \tag{28}$$

where $\beta_{\text{primal}} = \{\mathbb{I}(T = t)/p(T \mid Z_1, Z_2, C_1, C_2)\} \times Y$. The above functional can be estimated by following a similar strategy discussed for the functional in Eq. 24.

In this section, we focused on providing a representation of the efficient influence function for $\psi(t)$ when the treatment is primal fixable and the model is an mb-shielded ADMG, and propose a naive estimation procedure based on this representation. A more detailed investigation of the local asymptotic efficiency behaviors is left for future work.

6. Estimation of Any Identified Target Parameter

Thus far we have discussed inference of the target $\psi(t)$ in a broad class of ADMGs defined by the primal fixability criterion. However, in arbitrary hidden variable causal models, $\psi(t)$ may be identified even if the treatment T is not p-fixable. The resulting identifying functional is given by truncated factorization of the nested Markov model introduced in Section 2.2.2. This strategy for identification of the target is known to be sound and complete (Richardson et al., 2017). That is, identification of the target parameter $\psi(t)$ in a hidden variable causal model associated with a DAG $\mathcal{G}(V \cup H)$ may be rephrased, without loss of generality, using its corresponding latent projection ADMG $\mathcal{G}(V)$. Specifically, for $Y^* \equiv \text{an}_{\mathcal{G}_{V \setminus T}}(Y)$,

$$\psi(t) = \sum_{Y^*} Y \times \prod_{D \in \mathcal{D}(\mathcal{G}_{Y^*})} \phi_{V \setminus D}(p(V); \mathcal{G}(V)) \Big|_{T=t}, \qquad \textit{(Truncated nested Markov factorization)} \tag{29}$$

Algorithm 2 Nested IPW Functional $(\mathcal{G}(V), p(V), \tau)$

1: Let $Y^* = \text{an}_{\mathcal{G}_{V\setminus T}}(Y)$ and $D_T = \text{dis}_{\mathcal{G}}(T)$ and $\mathcal{D}^* \leftarrow \{D \in \mathcal{D}(\mathcal{G}_{Y^*}) \mid D \cap D_T \neq \emptyset\}$
2: **if** $\exists D \in \mathcal{D}^*$ such that D is not intrinsic in $\mathcal{G}$ **then**
3: **return** Fail
4: Define $q_D(D \mid \text{pa}_{\mathcal{G}}(D)) \equiv \phi_{V\setminus D}(p(V); \mathcal{G}(V))$
5: $\beta_{\text{nested}} \equiv \dfrac{\mathbb{I}(T=t)}{p(T \mid \text{mp}_{\mathcal{G}}(T))} \times \prod_{D\in\mathcal{D}^*} \Big(\dfrac{q_D(D \mid \text{pa}_{\mathcal{G}}(D))}{\prod_{D_i \in D} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))}\Big) \times Y$
6: **return** $\psi(t)_{\text{nested}} \equiv \mathbb{E}[\beta_{\text{nested}}]$

provided every $D \in \mathcal{D}(\mathcal{G}_{Y^*})$ is intrinsic in $\mathcal{G}(V)$; otherwise, $\psi(t)$ is not identifiable (Richardson et al., 2017). Recall from the definition of the nested Markov model that a set $S \subseteq V$ is said to be intrinsic in $\mathcal{G}$ if $V \setminus S$ is fixable, and $\phi_{V\setminus S}(\mathcal{G})$ contains a single district.

In special cases, when all observed variables are either discrete or multivariate normal, a parametric likelihood can be specified for the nested Markov model (Shpitser et al., 2018; Evans and Richardson, 2019), which leads naturally to estimation of $\psi(t)$ in Eq. 29 by the plug-in principle. However, assuming a full parametric likelihood is often unrealistic. Here we describe estimators that use only subsets of the likelihood thus reducing the chance of model misspecification.

6.1 Nested IPW Estimators

In Algorithm 2, we describe a generalization of the primal IPW estimator, introduced in Section 4, for *any* $\psi(t)$ that is identifiable from the observed margin $p(V)$ corresponding to an ADMG $\mathcal{G}(V)$. As these estimators are derived from the nested Markov factorization of the latent projection ADMG $\mathcal{G}(V)$, we coin the term *nested IPW* in referring to them.

Algorithm 2 takes as inputs the observed margin $p(V)$, the corresponding ADMG $\mathcal{G}(V)$, and a valid topological order τ; the algorithm then proceeds as follows. In line 1 it identifies districts D in $\mathcal{G}_{Y^*}$ such that $q_D(D \mid \text{pa}_{\mathcal{G}}(D))$ does not appear in the district factorization of the original ADMG $\mathcal{G}(V)$. We denote this set of districts by $\mathcal{D}^*$ and note that this set corresponds to districts in $\mathcal{G}^*_Y$ that have some intersection with D_T (the district of the treatment in $\mathcal{G}$.) We show that $\psi(t)$ is identifiable under the assumptions implied by $\mathcal{G}(V)$ if and only if each district in $D \in \mathcal{D}^*$ is intrinsic in $\mathcal{G}$; the algorithm checks this criterion in line 2, and returns "Fail" when it is not satisfied (corresponding to a non-identified target.) When $\psi(t)$ is identified, the nested IPW functional created in line 5 of the algorithm can be viewed as modifications to the district factorization of the ADMG $\mathcal{G}(V)$ involving the replacement of pieces of the kernel $q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))$ with the relevant intrinsic kernels that recreate the truncated nested Markov factorization in Eq. 29 (note there is no replacement intrinsic kernel involving T as a random variable, hence the truncation.) Indeed, we show the equivalence of this nested IPW functional and the truncated nested Markov factorization in the proof of Theorem 14 in Appendix H.

Theorem 14 formalizes that the nested IPW algorithm is sound and complete. That is, when Algorithm 2 returns a nested IPW functional, $\psi(t)_{\text{nested}} = \psi(t)$ and when the algorithm fails to return a functional, $\psi(t)$ is not identifiable within the given model.

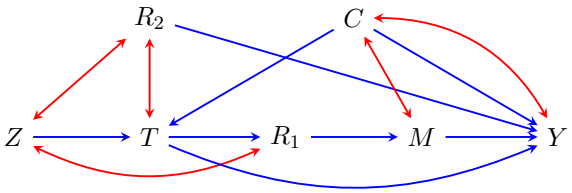


Figure 5: An ADMG where the treatment is not p-fixable but $\psi(t)$ is still identified via the truncated nested Markov factorization.

Theorem 14 (Soundness and completeness of Algorithm 2)
Let $p(V)$ and $\mathcal{G}(V)$ be the observed marginal distribution and ADMG induced by a hidden variable causal model associated with a DAG $\mathcal{G}(V \cup H)$. Then if $\psi(t)$ is identifiable in the model, $\psi(t) = \psi(t)_{nested}$ in the nested Markov model associated with $\mathcal{G}(V)$. If $\psi(t)$ is not identifiable in the model, Algorithm 2 returns 'fail'.

Note that $\psi(t)$ and $\psi(t)_{\text{nested}}$ are only equal in the model associated with $\mathcal{G}(V)$. In other words, equality of these two functionals may depend on equality restrictions implied by the model, and the two functionals may not be equal in an unrestricted observed data model.

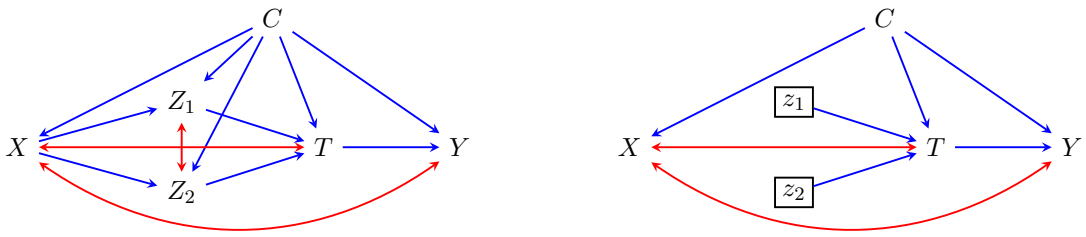
Nested IPW estimators are obtained by the empirical evaluation of the outer expectation in line 6 and the plug-in principles. As the above algorithm suggests, such estimators rely on specifying *only* a subset of the nested Markov likelihood that form the district of T. If all variables in D_T are discrete, this can be done using a (conditional) Moebius parameterization as described in Evans and Richardson (2019). In general, for estimating $\psi(t)_{\text{nested}}$ we rely on the correct specification of the kernel $q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T)) \equiv \phi_{V \setminus D_T}(p(V); \mathcal{G}(V))$ as follows. Given the topological ADMG factorization in (7), we can write down $q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T)) = \prod_{V_i \in D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))$. Each kernel $q_D(D \mid \text{pa}_{\mathcal{G}}(D))$ for $D \in \mathcal{D}^*$ can then be obtained via conditioning, marginalization, or fixing operations on the top-level kernel $q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))$. Hence, in addition to some regularity constraints and positivity assumptions, the nested IPW remains consistent if all the conditional densities in the set $p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)), \forall V_i \in D_T$ are correctly specified.

6.1.1 Example: A Nested IPW Functional

Consider the ADMG in Fig. 5. In this graph, T is not p-fixable. However, $\psi(t)$ is still identifiable via the truncated nested Markov factorization as follows. $Y^* = \{Y, M, C, R_1, R_2\}$ and $\mathcal{D}(\mathcal{G}_{Y^*}) = \{\{Y, M, C\}, \{R_1\}, \{R_2\}\}$. Fix a valid topological order $\tau = R_2 \prec Z \prec C \prec T \prec R_1 \prec M \prec Y$. Then from Eq. 29 we have,

$$\begin{aligned}\psi(t) &= \sum_{Y^*} Y \times \phi_{V \setminus \{Y,M,C\}}(p(V); \mathcal{G}) \times \phi_{V \setminus R_1}(p(V); \mathcal{G}) \times \phi_{V \setminus R_2}(p(V); \mathcal{G}) \Big|_{T=t} \\ &= \sum_{C,R_1,R_2,M} p(C) \times p(M \mid C, R_1) \times \mathbb{E}[Y \mid M, C, R_1, R_2, t] \times \sum_{Z} p(Z) \times p(R_1 \mid t, Z) \times p(R_2). \quad (30)\end{aligned}$$

We use Algorithm 2 to find another alternative for estimating the target $\psi(t)$. Note that $\mathcal{D}^*$ simply focuses on the districts related to $\mathcal{G}_{Y^*}$ that do not overlap with D_T. Therefore,



(a) $\mathcal{G}(V)$ (b) $\phi_Z(\mathcal{G}(V))$

Figure 6: (a) An ADMG where the treatment is not p-fixable but $\psi(t)$ is still identified via the truncated nested Markov factorization. (b) A valid sequence of fixing that yields a CADMG where T is p-fixable and $p(Y(t))$ can be obtained as $p(Y(t, z_1, z_2))$

$\mathcal{D}^*$ in line 1 of the algorithm is $\{\{R_1\}, \{R_2\}\}$. Since both of these districts are intrinsic in $\mathcal{G}$, Algorithm 2 does not fail, and we get

$$\psi(t)_{\text{nested}} = \mathbb{E}\Bigg[\frac{\mathbb{I}(T=t)}{p(T \mid R_2, Z, C)} \times \frac{\sum_Z p(Z) \times p(R_1 \mid T, Z)}{p(R_1 \mid T, Z, C, R_2)} \times \frac{p(R_2)}{p(R_2)} \overset{1}{\nearrow} \times Y \Bigg]. \tag{31}$$

At first glance, the models for $p(R_1 \mid T, Z)$ and $p(R_1 \mid T, Z, C, R_2)$ might look incompatible. However, the topological factorization of q_{D_T} provides us with a congenial representation in terms of specifying the conditional densities $p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)), \forall V_i \in D_T$. As mentioned earlier, if variables in the district of T can be assumed to be discrete, the Moebius parameterizaton (Evans and Richardson, 2019) may also be used to parameterize q_{D_T}.

7. Alternative Strategies When the Treatment is not Primal Fixable

We close our theoretical discussions with an example of ADMGs where T is not p-fixable (but the effect is still identifiable) and V can be partitioned as follows: baseline confounders C, the district of T denoted by D_T which includes the outcome Y, and instrumental variables Z that affect the outcome only via the treatment. The ADMG in Fig. 6(a) is an example. Besides having the nested IPW to estimate the effect, we make further progress in setting up estimating equations that exhibit partial robustness to model misspecification.

Consider the ADMG in Fig. 6(a) with $C = \{C\}, D_T = \{T, X, Y\}$, and $Z = \{Z_1, Z_2.\}$ The joint distribution district factorizes as follows,

$$p(V) = q_C(C) \times q_Z(Z \mid \text{pa}_{\mathcal{G}}(Z)) \times q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T)), \tag{32}$$

where

$$\begin{aligned} q_C(C) &= p(C), \\ q_Z(Z \mid \text{pa}_{\mathcal{G}}(Z)) &= p(Z \mid C, X), \text{ and} \\ q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T)) &= p(X \mid C) \times p(T, Y \mid C, X, Z). \end{aligned}$$

The treatment T is not primal fixable as it does not meet the condition that $\text{dis}_{\mathcal{G}}(T) \cap \text{ch}_{\mathcal{G}}(T) = \emptyset$. However, the target $\psi(t)$ is indeed identified via truncated nested Markov

factorization as follows. $Y^* = \{C, Y\}$ and $\mathcal{D}(\mathcal{G}_{Y^*}) = \{\{Y\}, \{C\}\}$.

$$\begin{aligned}\psi(t) &= \sum_{Y,C} Y \times \phi_{V\setminus Y}(p(V); \mathcal{G}) \times \phi_{V\setminus C}(p(V); \mathcal{G})\Big|_{T=t} \\ &= \sum_{Y,C} Y \times q_Y(Y \mid T = t, C, X, Z) \times q_C(C),\end{aligned}$$

where $q_C(C) = p(C)$ and

$$q_Y(Y \mid T, C, X, Z) = \frac{\sum_X q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))}{\sum_{X,Y} q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))} = \frac{\sum_X p(X \mid C) \times p(T \mid C, X, Z) \times p(Y \mid T, C, X, Z)}{\sum_{X,Y} p(X \mid C) \times p(T \mid C, X, Z) \times p(Y \mid T, C, X, Z)}.$$

Besides the nested IPW, we can estimate $\psi(t)$ by solving the estimating equation $\mathbb{P}_n[U^q_{\psi_t}] = 0$, where

$$U^q_{\psi_t} = \frac{\mathbb{I}(T = t) \times \Big(Y - \mathbb{E}_{q_Y}\big[Y \mid T = t, C, X, Z\big]\Big)}{q_Z(Z \mid \text{pa}_{\mathcal{G}}(Z)) \times q_{D_T\setminus Y}(D_T \setminus Y \mid \text{pa}_{\mathcal{G}}(D_T))} + \mathbb{E}_{q_Y}\big[Y \mid T = t, C, X, Z\big] - \psi(t)$$

and $\mathbb{E}_{q_Y}[Y \mid T, C, X, Z] = \sum_Y Y \times q_Y(Y \mid T, C, X, Z)$ and $q_Z(Z \mid \text{pa}_{\mathcal{G}}(Z)) = p(Z \mid \text{pa}_{\mathcal{G}}(Z))$. The resulting estimator resembles AIPW and exhibits a partial double robustness property; albeit using kernels that form pieces of the nested Markov likelihood. Upon correct specification of $p(X \mid C)$ and $p(T \mid C, X, Z)$, the obtained estimator is doubly robust in correct specification of either $p(Z \mid C, X)$ or $\mathbb{E}[Y \mid T, C, X, Z]$.

The above example suggests the following question: is it always possible to find an adjustment set in a kernel obtained after a sequence of valid fixing that would yield an estimating equation with partial robustness behaviours in the obtained estimator. In the types of ADMGs we considered in this subsection, we can always fix the instrumental variables and find a valid adjustment set in the obtained CADMG. In general, the structure preventing the validity of covariate adjustment is an inducing backdoor path, i.e., a bidirected path between T and Y where every non-endpoint is an ancestor of T or Y. It is known that there exists no separating set to block such paths (Verma and Pearl, 1990).[5] As an example, it is easy to verify that no valid adjustment set exists for the effect of T on Y in n Fig 6(a), due to the presence of the inducing backdoor path $T \leftrightarrow X \leftrightarrow Y$. However, upon fixing the instrumental variables Z_1 and Z_2, the path $T \leftrightarrow X \leftrightarrow Y$ is no longer inducing and we can find an adjustment set. The resulting CADMG is shown in Fig. 6(b) which yields the post-intervention distribution $p(C, X, T(z), Y(z))$. In this CADMG the effect of T on Y is identified by adjusting for C.

Finally, it is worth noting that even though T is not p-fixable in Fig 6(a), it is p-fixable in a CADMG obtained after a valid sequence of p-fixing operation ($\{Z_1, Z_2, X\}$). This yields a general procedure for identification of the target parameter via a sequence of p-fixing operation. The details on this procedure are deferred to Appendix F, where we define the primal fixing operator that operationalizes primal IPW so that it can be recursively applied to simplify problems where the treatment is not directly p-fixable. This results in estimating equations that are sequentially reweighted and partially doubly robust in the final nuisance models used after reweighting.

5. Verma constraints are independences that occur when such paths are broken via fixing operations.

8. Simulated Data Analysis

In this section, we describe a set of simulations to illustrate the key results presented in this paper. For each experiment, we generate data according to hidden variable DAGs that give rise to the latent projection ADMGs used in the motivating examples throughout the paper. Specifically, for each bidirected edge in the latent projection ADMG, we allow for the presence of unmeasured confounders that are parents of both end points of the bidirected edge; they are sampled from either a normal distribution, a uniform distribution, and/or a Bernoulli distribution. For example in Fig. 2(b), for the bidirected edge $T \leftrightarrow L$ the underlying hidden variable DAG contains variables H_1, H_2, and H_3 which are parents of both T and L. We provide an example of a data generating process in Appendix G. We use generalized additive models to fit all of the nuisance models. Our form of model misspecification involves dropping some of the appropriate conditioning variables and interaction terms in the nuisance models to demonstrate robustness to arbitrary model misspecification. The R code is available upon request.[6]

We consider two examples where treatment is primal fixable and an example where treatment is not primal fixable but the effect is nonetheless nonparametrically identified. In primal cases, we generated data according to hidden variable DAGs $\mathcal{G}(V \cup H)$ that give rise to the latent projection ADMGs $\mathcal{G}(V)$ shown in Figures 2(b) and 3. In the case where treatment is not primal fixable, we generated data according to a hidden variable DAG $\mathcal{G}(V \cup H)$ that gives rise to the latent projection ADMG $\mathcal{G}(V)$ shown in Figure 5. We analyzed the bias, variance, and robustness behaviors of our proposed estimators (Primal IPW, Dual IPW, Augmented Primal IPW, and Nested IPW) and compared them with the plug-in estimators. We further, evaluated the performance of the efficient influence function in the mb-shielded ADMG of Figure 3.

Simulation 1. Bias behavior of Primal, Dual, and Augmented Primal IPW estimators

We evaluated the bias of the proposed estimators as a function of sample size for the causal effect of a binary treatment T on a continuous outcome Y in the ADMGs of Figures 2(b) and 3. The estimators for the corresponding counterfactual means are provided in Sections 4.2.1, 4.3.3, and 5.1. These evaluations are carried under four different statistical models: (i) all conditional pieces in $\mathcal{M}_{\mathbb{M}\cup\mathbb{L}}$ are correctly specified, (ii) all conditional pieces in $\mathcal{M}_{\mathbb{M}\cup\mathbb{L}}$ are misspecified, (iii) only the conditional pieces in $\mathcal{M}_{\mathbb{M}}$ (terms in β_{dual}) are misspecified, and (iv) only the conditional pieces in $\mathcal{M}_{\mathbb{L}}$ (terms in β_{primal}) are misspecified. Results are reported and compared to the plug-in estimators in Figure 7. The x-axis is the sample size n with a range $(200, 15000)$ with increments of 200. For a given sample size, we iterate over 100 replications and report the absolute value of the estimator's bias for the causal effect, averaged over all the iterations. The error bars illustrate deviations from the mean where the length equals the standard error of the mean. The theory is borne out by the simulations: in both ADMGs, the plug-in estimator is unbiased only under scenario (i); the primal IPW estimator is unbiased under scenarios (i) and (iii); the dual IPW estimator is unbiased under scenarios (i) and (iv); the APIPW estimator remains unbiased when at least one set of models corresponding to those used in β_{primal} or β_{dual} are correctly specified,

6. For Python implementations, see the open source package Ananke (link: `https://ananke.readthedocs.io/en/latest/`)

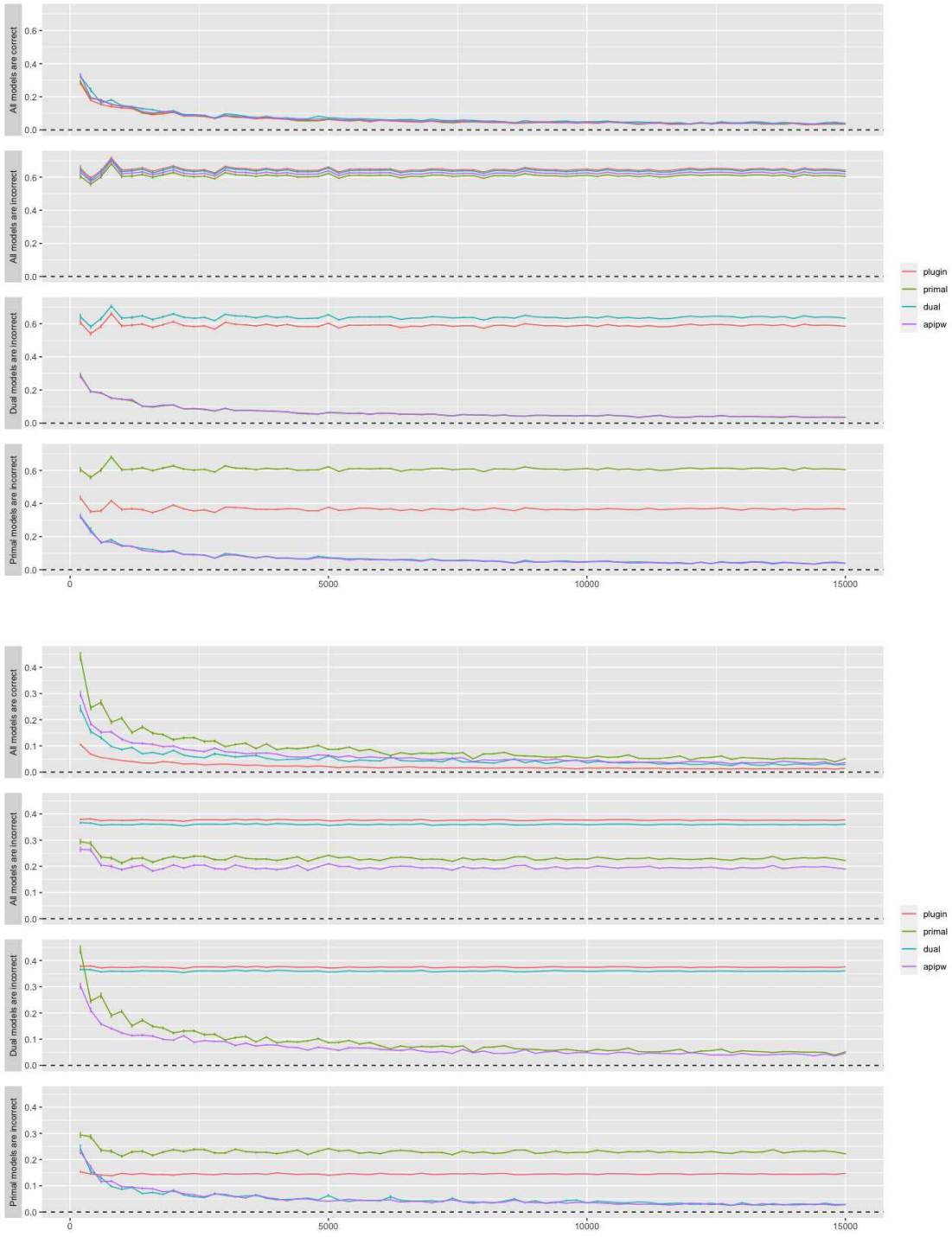


Figure 7: Bias behavior as a function of sample size when (1st row) all models are correctly specified, (2nd row) all models are incorrect, (3rd row) only dual models are misspecified, or (4th row) when only primal models are misspecified. The panel on (top) uses the ADMG in Fig. 2(b), and the (bottom) one uses the ADMG in Fig. 3. The error bars illustrate deviations across the multiple iterations.

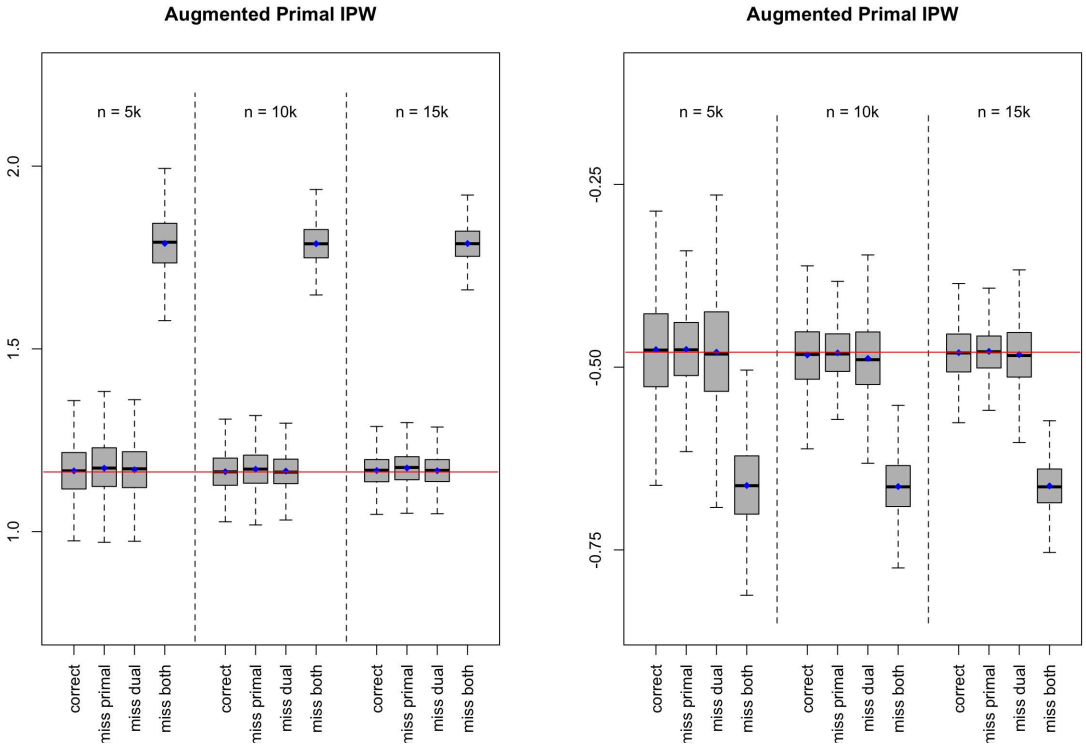


Figure 8: Demonstrating the double robustness of the Augmented Primal IPW estimator. The boxplot panel on the (left) uses the ADMG in Fig. 2(b), and the one on the (right) uses the ADMG in Fig. 3. The red dashed lines indicate the true values of the ACE.

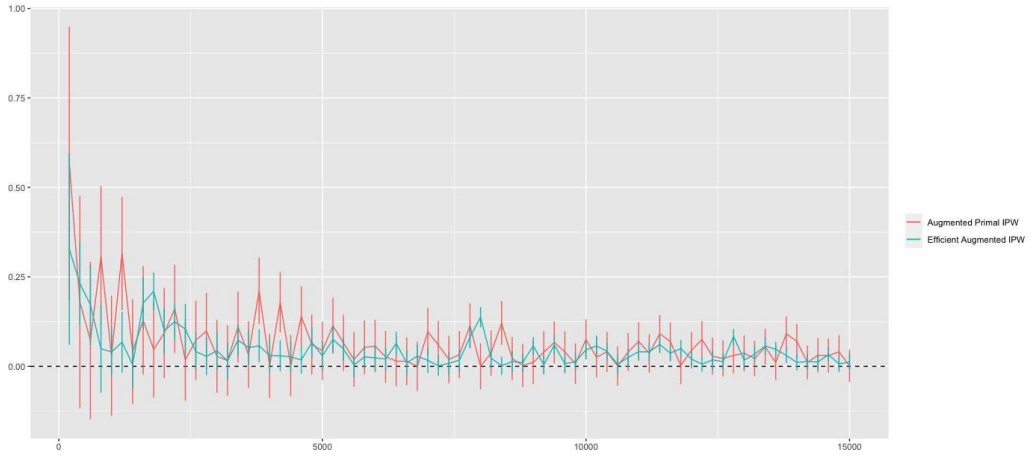


Figure 9: Comparing bias and variance behavior of the Augmented Primal IPW and efficient APIPW as a function of sample size, using the ADMG in Fig. 3. The error bars illustrate the variance of the corresponding influence function.

and is biased if they are all misspecified. That is, the influence function based estimator is unbiased in scenarios (i), (iii), and (iv). See Appendix G for more simulations.

Simulation 2. Double robustness of Augmented Primal IPW

To illustrate the double robustness behavior of the APIPW estimator, we provide the boxplots of the causal effect of T on Y in the ADMGs of Figures 2(b) and 3, under different settings of model misspecifications. The simulations are replicated 1000 times. For comparison, we plot the results for three sample sizes, $n = 5k, 10k$, and $15k$ in Figure 8. The true causal effect is 1.16 in Fig. 2(b) and is -0.48 in Fig. 3. The blue dot on each boxplot corresponds to the mean. The results can also be compared on the basis of means and standard errors. We provide a similar set of boxplots in Appendix G to highlight the standard deviations across different scenarios.

Simulation 3. Efficiency of Augmented Primal IPW in mb-shielded ADMGs

As pointed out in Section 5, we can exploit constraints in a statistical model of an ADMG that is not nonparametrically saturated to obtain more efficient estimators. We discussed the efficiency results for the special class of mb-shielded ADMGs. The ADMG in Figure 2(b) is nonparametrically saturated (no constraints). However, the ADMG in Figure 3 is not saturated (there are constraints) and it is an mb-shielded ADMG. Hence, we can obtain an estimator that is more efficient than the nonparametric influence function based estimator APIPW. The plots in Figure 9 summarize the simulation results comparing variances of the APIPW and the efficient influence function based estimators for the causal effect in the ADMG of Figure 3. The form of the efficient influence function is given in Section 5.1. The y-axis is the bias and the error bars illustrate the variance of the estimators (which is the variance of the corresponding influence functions). As expected, though both estimators are unbiased, the one based on the efficient influence function offers lower variance with varying sample sizes. The absolute reduction in variance is greater and most beneficial at smaller sample sizes, but the relative reduction in variance remains the same across all sample sizes. This highlights the benefit of using the efficient IF in many practical applications where sample sizes may be small.

Simulation 4. Bias behavior of Nested IPW estimator when T is not primal fixable

We evaluated the bias of the nested IPW estimator as a function of sample size for the causal effect of T on Y in the ADMG of Figure 5 where treatment T is not primal fixable. We compared this to the behavior of the plug-in estimator. These estimators are discussed in Section 6.1.1. The evaluations are carried under three different statistical models: (i) all conditional pieces in the district of T along with all pieces outside of the district of T are correctly specified, (ii) the outcome regression is misspecified (a term outside of the district of T), and (iii) the district of T is misspecified by misspecifying the treatment propensity. Results are reported and compared to the plug-in estimator in Figure 10. The x-axis is the sample size n with a range $(200, 15000)$ with increments of 200. For a given sample size, we iterate over 100 replications and report the absolute value of the bias for the causal effect, averaged across all iterations. The error bars illustrate deviations from the mean. and the length correspond to the standard error of the mean. The experiments support

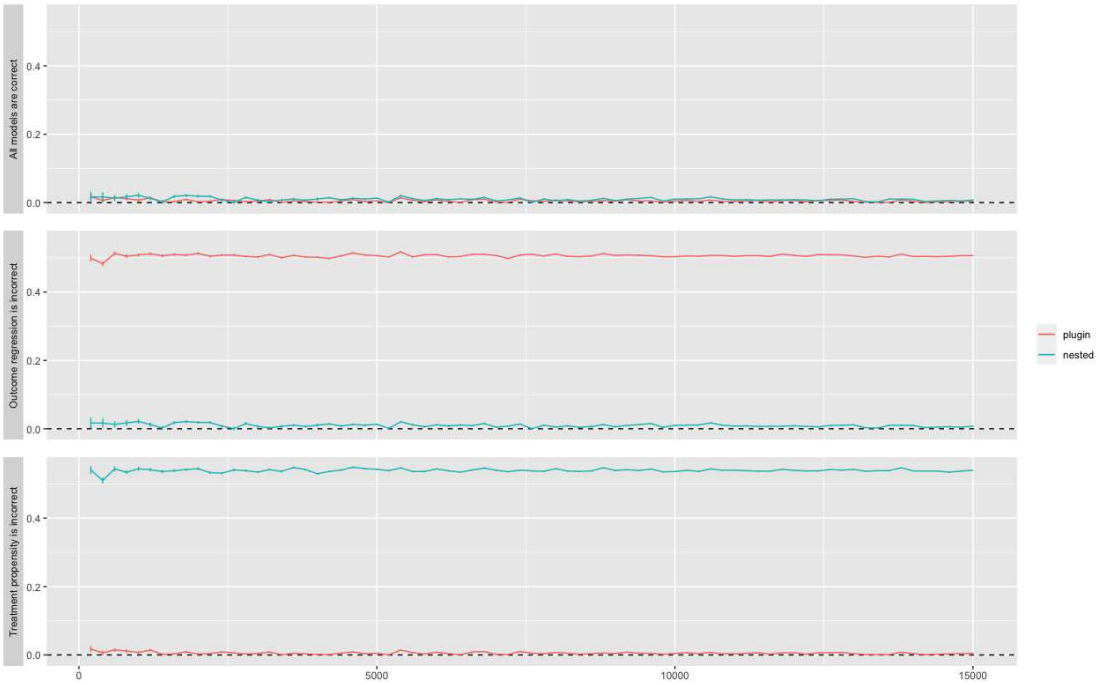


Figure 10: Bias behavior as a function of sample size when (1st row) all models are correctly specified, (2nd row) outcome regression is incorrect, and (3rd row) propensity score is incorrect, using the ADMG in Fig. 5. The error bars illustrate deviations across the multiple iterations.

our theory: the nested IPW estimator is biased under scenario (iii), but remains unbiased under scenarios (i) and (ii).

9. Conclusions

In this paper, we bridged the gap between identification and estimation theory for the causal effect of a single treatment on a single outcome in hidden variable causal models associated with directed acyclic graphs (DAGs). We provided a simple graphical criterion, primal fixability, which when satisfied allows for the derivation of two novel IPW estimators – primal and dual IPW. We further derived the nonparametric influence function under primal fixability of the treatment that yields the augmented primal IPW estimator and showed that it is doubly robust in the models used in primal and dual IPW estimators. We showed that in a strict subclass of primal fixability, when treatment is ordinary fixable, we can always find a valid adjustment set that permits the use of an influence function based estimator of augmented inverse probability weighted estimator to settings with hidden variables. We considered restrictions on the tangent space implied by the latent projection acyclic directed mixed graph (ADMG) of the hidden variable causal model. We provided an algorithm (Algorithm 1), that is sound and complete for the purposes of checking the nonparametric saturation status of a hidden variable causal model as long as these hidden

variables are unrestricted. Further, through the use of mb-shielded ADMGs, we provided a graphical criterion that defines a class of hidden variable causal models whose score restrictions resemble those of a DAG with no hidden variables. For the class of causal models that can be expressed as an mb-shielded ADMG, we then derived the form of the efficient influence function under primal fixability, that takes advantage of the Markov restrictions implied on the observed data. Finally, we developed a weighting estimation strategy for *any* identifiable causal effect involving a single treatment and a single outcome. We call these estimators *nested IPW* which only rely on the conditional densities involving variables in the district of the treatment. A natural extension of the present work is deriving influence function based estimators for any identifiable causal effect (including those that involve multiple treatment variables), and finding their most efficient versions by projecting onto the tangent space defined by equality restrictions, such as conditional independences and Verma constraints, implied by the causal model.

Appendices

A. Glossary of Terms and Notations

Symbol	Definition	Symbol	Definition
T	Treatment	$\mathcal{G}(V)$	Graph $\mathcal{G}$ with vertices V
$Y, Y(t)$	Outcome, potential outcome	$\mathcal{G}(V, W)$	A CADMG with fixed W
V	Observed variables	$\mathcal{G}_S$	Subgraph of $\mathcal{G}$ on vertices S
H	Unmeasured variables	$\text{pa}_{\mathcal{G}}(V_i)$	Parents of V_i in $\mathcal{G}$
W	Fixed variables	$\text{ch}_{\mathcal{G}}(V_i)$	Children of V_i in $\mathcal{G}$
$\psi(t)$	Target parameter $\mathbb{E}[Y(t)]$	$\text{an}_{\mathcal{G}}(V_i)$	Ancestors of V_i in $\mathcal{G}$
U_{ψ_t}	Influence function for $\psi(t)$	$\text{de}_{\mathcal{G}}(V_i)$	Descendants of V_i in $\mathcal{G}$
$\mathbb{H}$	Hilbert space	$\text{mb}_{\mathcal{G}}(V_i)$	Markov blanket of V_i in $\mathcal{G}$
Λ	Tangent space	$\text{mp}_{\mathcal{G}}(V_i)$	Markov pillow of V_i in $\mathcal{G}$
$\Lambda^{\perp}$	Orthogonal complement	$\text{dis}_{\mathcal{G}}(V_i)$	District of V_i in $\mathcal{G}$
$\mathcal{U}$	Class of all influence functions	D_T	District of T
U_{ψ}^{eff}	Efficient influence function	$\mathcal{D}(\mathcal{G})$	Set of all districts in $\mathcal{G}$
$\pi[h \mid \Lambda]$	Projection of h onto Λ	τ	A valid topological order
$\mathbb{C}$	Pre-treatment variables	$V_i \prec V_j$	V_i precedes V_j
$\mathbb{L}$	Post-treatment variables in D_T	$\{\prec V_i\}$	Vertices preceding V_i
$\mathbb{M}$	Variables not in $\mathbb{C} \cup \mathbb{L}$	$\phi_{V_i}(\mathcal{G})$	Fixing V_i in $\mathcal{G}$
$\mathbb{M}^*$	Inverse Markov pillow of T	$\phi_{V_i}(q_V; \mathcal{G})$	Fixing V_i in $q_V(\cdot \mid \cdot)$
$\mathbb{P}_n$	Empirical distribution	$\phi_{\neg S}(\mathcal{G})$	CADMG from fixing $V \setminus S$ recursively
$S(V)$	Score of $p(V)$	Y^*	$\text{an}_{\mathcal{G}_{V \setminus T}}(Y)$

B. Example of Latent Projection

Fig. 11(a) shows an illustrative example of a hidden variable DAG $\mathcal{G}(V \cup H)$ and the ADMG $\mathcal{G}(V)$ obtained by applying rules of latent projection described in Section 2.

(a) $\mathcal{G}(V \cup H)$ (b) $\mathcal{G}(V)$

Figure 11: (a) A hidden variable DAG $\mathcal{G}(V \cup H)$. (b) The ADMG $\mathcal{G}(V)$ obtained via latent projection.

C. Marginalization, Conditioning, and Fixing in Kernels

A kernel $q_V(V \mid W)$ is a mapping from values of W to normalized densities over V. That is, $\sum_V q_V(V \mid W = w) = 1, \forall w \in W$. For any set of variables $X \subseteq V$, marginalization and conditioning in a kernel are defined as follows.

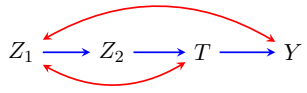
$$q_{V \setminus X}(V \setminus X \mid W) \equiv \sum_X q_V(V \mid W), \text{ and}$$
$$q_V(V \setminus X \mid X, W) \equiv \frac{q_V(V \mid W)}{q_V(X \mid W)}.$$

The notation $q_V(\cdot \mid X)$ makes clear which variables appearing past the "conditioning" bar in a kernel are fixed as opposed to simply conditioned on. That is, if a variable $X_i \notin V$, then it is fixed, else it is conditioned on. Occasionally, fixing operations may also simplify to marginalization or conditioning events. We illustrate these concepts with a simple example.

Consider the ADMG shown in Fig. 12(a) and fix the kernel of interest to be $q_Y(Y \mid T, Z_1, Z_2)$, i.e., a kernel where all other variables except Y are fixed. A valid fixing sequence in order to obtain such a kernel from the joint $p(V)$ is (Z_2, Z_1, T). Fixing Z_2 entails dividing by the simple conditional $p(Z_2 \mid Z_1)$ and yields the CADMG $\phi_{Z_2}(\mathcal{G})$ and corresponding kernel $q_{Z_1,T,Y}(Z_1, T, Y \mid Z_2)$ shown in Fig. 12(b). In order to fix Z_1, we must divide by the kernel $q_{Z_1,T,Y}(Z_1 \mid Z_2, T, Y)$. By rules of conditioning and marginalization in kernels,

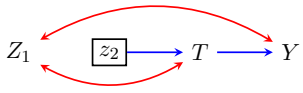
$$q_{Z_1,T,Y}(Z_1 \mid Z_2, T, Y) \equiv \frac{q_{Z_1,T,Y}(Z_1, T, Y \mid Z_2)}{q_{Z_1,T,Y}(T, Y \mid Z_2)} \equiv \frac{q_{Z_1,T,Y}(Z_1, T, Y \mid Z_2)}{\sum_{Z_1} q_{Z_1,T,Y}(Z_1, T, Y \mid Z_2)}$$

Fixing Z_1 and evaluating the above expression gives us the CADMG and corresponding kernel shown in Fig. 12(c). That is, fixing Z_1 in the kernel $q_{Z_1,T,Y}(Z_1 \mid Z_2, T, Y)$, simplifies to marginalization of Z_1. Finally, applying rules of conditioning and marginalization to the kernel $q_{T,Y}(T, Y \mid Z_1, Z_2)$ we can obtain the kernel $q_{T,Y}(T \mid Z_1, Z_2, Y)$. Dividing by this corresponds to fixing T, giving us the CADMG and desired kernel shown in Fig. 12(d).



(a) $\mathcal{G}$

$p(V) = p(Z_2 \mid Z_1) \times p(Z_1) \times p(T, Y \mid Z_1, Z_2)$



(b) $\phi_{Z_2}(\mathcal{G})$

$q_{Z_1,T,Y}(Z_1, T, Y \mid Z_2) = p(Z_1) \times p(T, Y \mid Z_1, Z_2)$

(c) $\phi_{\{Z_1,Z_2\}}(\mathcal{G})$

$q_{T,Y}(T, Y \mid Z_1, Z_2) = \sum_{Z_1} p(Z_1) \times p(T, Y \mid Z_1, Z_2)$

(d) $\phi_{\{Z_1,Z_2,T\}}(\mathcal{G})$

$q_Y(Y \mid T, Z_1, Z_2) = \frac{\sum_{Z_1} p(Z_1) \times p(T, Y \mid Z_1, Z_2)}{\sum_{Z_1} p(Z_1) \times p(T \mid Z_1, Z_2)}$

Figure 12: An example to illustrate fixing and kernel operations.

D. An Overview of Semiparametric Estimation Theory

Assume a statistical model $\mathcal{M} = \{p_\eta(Z) : \eta \in \Gamma\}$ where Γ is the parameter space and η is the parameter indexing a specific model. We are often interested in a function $\psi : \eta \in \Gamma \mapsto \psi(\eta) \in \mathbb{R}$; i.e., a parameter that maps the distribution P_η to a scalar number in $\mathbb{R}$, such as an identified average causal effect. (For brevity, we sometimes use ψ instead of $\psi(\eta)$, which should be obvious from context.) Truth is denoted by P_{η_0} and ψ_0. An estimator $\widehat{\psi}_n$ of a scalar[7] parameter ψ based on n i.i.d copies $Z_1, \ldots, Z_n$ drawn from $p_\eta(Z)$, is *asymptotically linear* if there exists a measurable random function $U_\psi(Z)$ with mean zero and finite variance such that

$$\sqrt{n} \times (\widehat{\psi}_n - \psi) = \frac{1}{\sqrt{n}} \times \sum_{i=1}^{n} U_\psi(Z_i) + o_p(1), \tag{33}$$

where $o_p(1)$ is a term that converges in probability to zero as n goes to infinity. The random variable $U_\psi(Z)$ is called the *influence function* of the estimator $\widehat{\psi}_n$. The term influence function comes from the robustness literature (Hampel, 1974).

Before mentioning the asymptotic properties of an asymptotically linear estimator, it is worth noting that in asymptotic theory, we can sometimes construct *super efficient* estimators, e.g. Hodges estimator, that have undesirable local properties associated with them. Therefore, the analysis is oftentimes restricted to *regular*[8] and asymptotically linear (RAL) estimators to avoid such complications. Although most reasonable estimators are RAL, regular estimators do exist that are not asymptotically linear. However, as a consequence of Hájek (1970) representation theorem, the most efficient regular estimator is asymptotically linear; hence, it is reasonable to restrict attention to RAL estimators. According to Newey (1990), the influence function of a RAL estimator is the same as the influence function of its estimand. Further, there is a bijective correspondence between RAL estimators and influence functions.

By a simple consequence of the central limit theorem and Slutsky's theorem, it is straightforward to show that the RAL estimator $\widehat{\psi}_n$ is *consistent and asymptotically normal* (CAN), with asymptotic variance equal to the variance of its influence function U_ψ,

$$\sqrt{n} \times (\widehat{\psi}_n - \psi) \xrightarrow{d} N\big(0,\ \mathrm{var}(U_\psi)\big). \tag{34}$$

The first step in dealing with a semiparametric model, is to consider a simpler finite-dimensional parametric submodel that is contained within the semiparametric model and it contains the truth. Consider a (regular) parametric submodel $\mathcal{M}_{\text{sub}} = \{P_{\eta_\kappa} : \kappa \in [0, 1)$ where $P_{\eta_{\kappa=0}} = P_{\eta_0}\}$ of the model $\mathcal{M}$. Given P_{η_0}, define the corresponding score to be $S_{\eta_0}(Z) = \dfrac{d}{d\kappa} \log p_{\eta_\kappa}(Z)\Big|_{\kappa=0}$. It is known that

$$\frac{d}{d\kappa}\psi(\eta_\kappa)\Big|_{\kappa=0} = \mathbb{E}\Big[U_\psi(Z) \times S_{\eta_0}(Z)\Big], \tag{35}$$

7. Here, our focus is on estimation of $\psi = \mathbb{E}[Y(t)]$ which is a scalar parameter. For an extension to a vector valued functional in $\mathbb{R}^q, q > 1$, refer to Tsiatis (2007); Bickel et al. (1993).

8. Given a collection of probability laws $\mathcal{M}$, an estimator $\widehat{\psi}$ of $\psi(P)$ is said to be regular in $\mathcal{M}$ at P if its convergence to $\psi(P)$ is locally uniform (van der Vaart, 2000).

where $\psi(\eta_\kappa)$ is the target parameter in the parametric submodel, $U_\psi(Z)$ is the corresponding influence function evaluated at law P_{η_0}, $S_{\eta_0}(Z)$ is the score of the law P_{η_0}, and the expectation is taken with respect to P_{η_0}. Equation 35 provides an easy way to derive an influence function for the parameter ψ. In the next subsection, we use this equation to derive an influence function for our target $\psi = \mathbb{E}[Y(t)]$ and discuss its properties.

Influence functions provide a geometric view of the behavior of RAL estimators. Consider a Hilbert space[9] $\mathbb{H}$ of all mean-zero scalar functions, equipped with an inner product defined as $\mathbb{E}[h_1 \times h_2], h_1, h_2 \in \mathbb{H}$. The *tangent space* in the model $\mathcal{M}$, denoted by Λ, is defined to be the mean-square closure of parametric submodel tangent spaces, where a parametric submodel tangent space is the set of elements $\Lambda_{\eta_\kappa} = \{\alpha S_{\eta_\kappa}(Z)\}, \alpha$ is a constant and S_{η_κ} is the score for the parameter ψ_{η_κ} for some parametric submodel. In mathematical form, $\Lambda = \overline{[\Lambda_{\eta_\kappa}]}$.

The tangent space Λ is a closed linear subspace of the Hilbert space $\mathbb{H}$ ($\Lambda \subseteq \mathbb{H}$). The orthogonal complement of the tangent space, denoted by $\Lambda^\perp$, is defined as $\Lambda^\perp = \{h \in \mathbb{H} \mid \mathbb{E}[h \times h'] = 0, \forall h' \in \Lambda\}$. Note that $\mathbb{H} = \Lambda \oplus \Lambda^\perp$, where $\oplus$ is the direct sum, and $\Lambda \cap \Lambda^\perp = \{0\}$. Given an arbitrary element $h \in \Lambda^\perp$, it holds that for any submodel $\mathcal{M}_{\text{sub}}$, with score S_{η_0} corresponding to P_{η_0}, $\mathbb{E}[h \times S_{\eta_0}] = 0$. Consequently, using Eq. 35, $h + U_\psi(Z)$ is also an influence function. The vector space $\Lambda^\perp$ is then of particular importance because we can now construct the class of all influence functions, denoted by $\mathcal{U}$, as $\mathcal{U} = U_\psi(Z) + \Lambda^\perp$. Upon knowing a single IF $U_\psi(Z)$ and the tangent space orthogonal complement $\Lambda^\perp$, we can obtain the class of all possible RAL estimators that admit the CAN property.

Out of all the influence functions in $\mathcal{U}$ there exists a unique one which lies in the tangent space Λ, and which yields the most efficient RAL estimator by recovering the *semiparametric efficiency bound*. This efficient influence function can be obtained by projecting any influence function, call it U^*_ψ, onto the tangent space Λ. This operation is denoted by $U^{\text{eff}_\psi} = \pi[U^*_\psi \mid \Lambda]$, where U^{eff}_ψ denotes the efficient IF.

On the other hand, if the tangent space contains the entire Hilbert space, i.e., $\Lambda = \mathbb{H}$, then the statistical model $\mathcal{M}$ is called a *nonparametric* model. In a nonparametric model, we only have one influence function since $\Lambda^\perp = \{0\}$. This unique influence function can be obtained via Eq. 35 and corresponds to the efficient influence function U^{eff}_ψ (the unique element in the tangent space Λ) in the nonparametric model $\mathcal{M}$. For a detailed description of the concepts outlined here, please refer to Tsiatis (2007); Bickel et al. (1993).

D.1 An Overview of Inference for the Adjustment Functional

Having briefly discussed causal models of a DAG in Section 2, we now provide a short overview of estimation theory surrounding the target $\psi(t)$ in such a model.

If a parametric likelihood can be correctly specified for the statistical DAG model of the observed data distribution, then an efficient estimator for $\psi(t)$ may be derived using the plug-in principle. In the commonly assumed case where the DAG corresponding to the observed data distribution is complete, the plug-in estimator for $\psi(t)$ reduces to $\mathbb{P}_n[\mu_t(C; \widehat{\eta_1})]$, where $\mathbb{P}_n[.] \coloneqq \frac{1}{n}\sum_{i=1}^n(.)$, $\mu_t(C; \eta_1)$ is the correctly specified parametric form for $\mathbb{E}[Y \mid T = t, C]$, and $\widehat{\eta_1}$ are the maximum likelihood values of η_1.

9. The Hilbert space of all mean-zero scalar functions is the L^2 space. For a precise definition of Hilbert spaces see Luenberger (1997).

Since assuming a correctly specified parametric observed data likelihood, or even a correctly specified outcome regression $\mu_t(C;\eta)$ is unrealistic in practice, a variety of other estimators have been developed that place *semiparametric* restrictions on the observed data distribution. One such estimator, based on *inverse probability weighting (IPW)*, seeks to compensate for a biased treatment assignment by reweighing observed outcomes of units assigned $T = t$ by the inverse of the normalized treatment assignment probability $p(T = t \mid C)$. If this probability has a known parametric form $\pi_t(C;\eta_2) \equiv p(T = t \mid C)$, the IPW estimator takes the form $\mathbb{P}_n[\frac{\mathbb{I}(T=t)}{\pi_t(C;\widehat{\eta_2})} \times Y]$, where $\mathbb{I}(.)$ is the indicator function, and $\widehat{\eta_2}$ are the maximum likelihood estimates of η_2. While the IPW estimator is inefficient, it is simple to implement, and is often used in cases where the treatment assignment model $\pi_t(C;\eta_2)$ is known by design, as is often the case in controlled trials.

The plug-in and IPW estimators of $\psi(t)$ are both *$\sqrt{n}$-consistent and asymptotically normal* if the models they rely on, $\mu_t(C;\eta_1)$ and $\pi_t(C;\eta_2)$ respectively, are parametric and correctly specified. Otherwise, these estimators are no longer consistent. If flexible models are used for $\mu_t(C)$ and $\pi_t(C)$ instead, the resulting estimators may remain consistent, but converge to the true value of $\psi(t)$ at unacceptably slow rates; see Chernozhukov et al. (2018) for examples.

A principled alternative is to consider influence functions and RAL estimators. In the nonparametric saturated model, corresponding to the complete DAG, the unique influence function for $\psi(t)$ is given by $U_{\psi_t} = \frac{\mathbb{I}(T=t)}{\pi_t(C)} \times \{Y - \mu_t(C)\} + \mu_t(C) - \psi(t)$, yielding the *AIPW* estimator: $\mathbb{P}_n\Big[\frac{\mathbb{I}(T=t)}{\pi_t(C;\widehat{\eta_2})} \times \{Y - \mu_t(C;\widehat{\eta_1})\} + \mu_t(C;\widehat{\eta_1})\Big]$. Given the standard factorization of the complete DAG as $p(Y \mid A, C) \times p(A \mid C) \times p(C)$, the propensity score model $\pi_t(C)$ and the outcome regression model $\mu_t(C)$ are variationally independent. Further, the bias of this estimator is a product of the biases of its nuisance functions $\pi_t(C)$ and $\mu_t(C)$. As a result, the AIPW estimator exhibits the *double robustness property*, where it remains consistent if *either* of the two nuisance models $\pi_t(C)$ or $\mu_t(C)$ is specified correctly, even if the other is arbitrarily misspecified.

In a semiparametric model of a DAG, which is defined by conditional independence restrictions on the tangent space implied by the DAG factorization, the above influence function can be projected onto the tangent space of the model to improve efficiency; see Rotnitzky and Smucler (2020) for details.

E. Intuitions for APIPW in the Nonparametric Model

Given a post treatment variable V_i and its conditional density $p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))$ in the identified functional of $\psi(t)$ in Eq. 14, there is a corresponding term in the influence function U_{ψ_t} in Theorem 8 of the form

$$f_1(\prec V_i) \times \Big(f_2(\preceq V_i) - \sum_{V_i} f_2(\preceq V_i) \times p\big(V_i \mid \text{mp}_{\mathcal{G}}(V_i)\big) \Big), \tag{36}$$

where $f_1(\prec V_i)$ denotes a function of variables that come before V_i in the topological order, a.k.a history/past of V_i. Similarly, $f_2(\preceq V_i)$ is a function of past of V_i and including V_i itself. $f_1(\prec V_i)$ is defined as follows,

$$f_1(\prec V_i) = \begin{cases} \dfrac{\mathbb{I}(T=t)}{\prod_{L_i \prec V_i} \; p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))}, & \text{if } V_i \in \mathbb{M} \\ \dfrac{\prod_{M_i \prec V_i} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec V_i} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))}, & \text{if } V_i \in \mathbb{L} \end{cases} \tag{37}$$

Interestingly, these weights resemble the ones in ψ_{primal} and ψ_{dual} that we introduced in Lemmas 4 and 5, if the target were the counterfactual mean $\mathbb{E}[V_i(t)]$. That is,

$$\psi_{v_i,\text{primal}} = \mathbb{E}\Big[\frac{\mathbb{I}(T=t)}{\prod_{L_i \prec V_i} \; p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \sum_T \prod_{L_i \prec V_i} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times V_i \Big]$$
$$\psi_{v_i,\text{dual}} = \mathbb{E}\Big[\frac{\prod_{M_i \prec V_i} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec V_i} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times V_i \Big].$$

However, to calculate the effect of T on Y, we do not need to worry about the effect of T on intermediate variables V_i. In Lemma 9, we show that the influence function U_{ψ_t} in Theorem 8 cleverly uses the information in these intermediate primal and dual estimators and yields a doubly robust estimator for ψ_t.

F. Primal Fixing Operator

In this section we introduce the primal fixing operator, which is a generalization of the fixing operator used in the definition of the nested Markov model. We show how a valid sequence of primal fixing serves as a useful identification strategy whose complexity lies between a single use of the primal fixing criterion (as seen in Section 3) and truncated nested Markov factorization (as seen in Section 6.) Proofs are deferred to Appendix H.

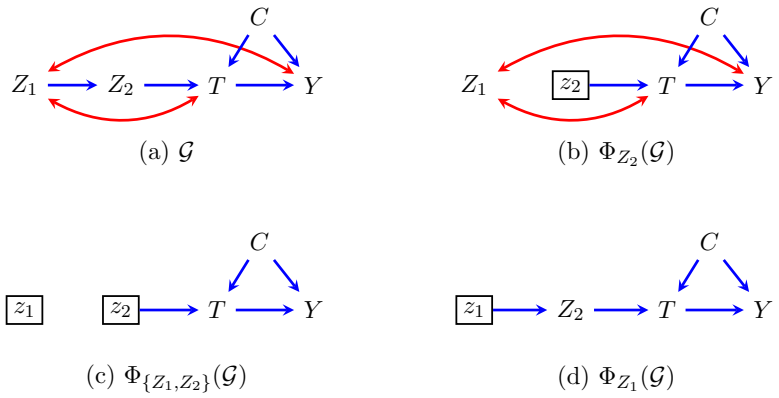
Given a kernel $q_V(V \mid W)$ that is nested Markov with respect to a CADMG $\mathcal{G}(V, W)$, we define a vertex V_i to be primal fixable (p-fixable) if it has no bidirected path to any of its children, i.e., $\text{dis}_{\mathcal{G}}(V_i) \cap \text{ch}_{\mathcal{G}}(V_i) = \emptyset$. If V_i is primal fixable in $\mathcal{G}(V, W)$, the graphical operation of primal fixing V_i applied to $\mathcal{G}$, denoted by $\Phi_{V_i}(\mathcal{G})$, yields a new CADMG $\mathcal{G}(V \setminus V_i, W \cup V_i)$ where all incoming edges into V_i are removed and V_i is fixed to some value v_i. That is, the graphical operation of primal fixing is exactly the same as ordinary fixing. However, the two forms of fixing differ in the definition of the probabilistic operators; the probabilistic operator for primal fixing is defined as follows. Given a kernel $q_V(V \mid W)$ that nested factorizes with respect to a CADMG $\mathcal{G}(V, W)$ in which V_i is primal fixable, let D_{V_i} denote the district of V_i in $\mathcal{G}(V, W)$. Then the probabilistic operation of primal fixing V_i denoted by $\Phi_{V_i}(q_V; \mathcal{G})$ is given by,

$$\begin{aligned}\Phi_{V_i}(q_V; \mathcal{G}) \equiv q_{V \setminus T}(V \setminus V_i \mid W \cup V_i) &\equiv \frac{q_V(V \mid W)}{q_{D_{V_i}}(V_i \mid \text{mb}_{\mathcal{G}}(V_i), W)} \\ &= q_V(V \mid W) \times \frac{\sum_{V_i} \prod_{D_i \in D_{V_i}} q_V(D_i \mid \text{mp}_{\mathcal{G}}(D_i), W)}{\prod_{D_i \in D_{V_i}} q_V(D_i \mid \text{mp}_{\mathcal{G}}(D_i), W)}. \end{aligned} \tag{38}$$

The second equality follows from the application of algebraic manipulations to the kernel $q_{D_{V_i}}(V_i \mid \text{mb}_{\mathcal{G}}(V_i), W)$ in almost the exact same manner as the ones shown in the primal IPW formulation in the main draft (paragraph below Lemma 4.) In fact, when $q_V(V \mid W)$ is defined to be the observed margin $p(V)$ of a hidden variable causal DAG, it is easy to see that the primal fixing operator recovers the primal IPW formula for $p(V(v_i))$.

Similar to the ordinary fixing operator ϕ, the primal fixing operation can be applied to sequences of vertices $\sigma_S = (S_1, \ldots, S_p)$ in a set S provided this forms a valid p-fixing sequence for S in $\mathcal{G}(V, W)$. The sequence is valid if S_1 is p-fixable in $\mathcal{G}$, S_2 is p-fixable in $\Phi_{S_1}(\mathcal{G})$, and so on. Given a sequence $\sigma_S = (S_1, \ldots, S_p)$ p-fixable in $\mathcal{G}(V, W)$, define $\Phi_{\sigma_S}(\mathcal{G}(V, W))$ as $\mathcal{G}$ if S is empty, and $\Phi_{\sigma_S \setminus S_1}(\Phi_{S_1}(\mathcal{G}))$ otherwise. Similarly, define $\Phi_{\sigma_S}(q_V; \mathcal{G})$ to be $q_V(V \mid W)$ if S is empty and $\Phi_{\sigma_S \setminus S_1}(\Phi_{S_1}(q_V; \mathcal{G}); \Phi_{S_1}(\mathcal{G}))$ otherwise. We say S is p-fixable in $\mathcal{G}(V, W)$ if there exists a valid sequence σ_S. Since the graphical operator of p-fixing is equivalent to ordinary fixing, it follows trivially that two valid p-fixing sequences yield the same CADMG. The following lemma formalizes that any two valid p-fixing sequences also yield the same kernel.

Lemma 15 (Commutativity of p-fixing)
If σ_S^1 and σ_S^2 are both valid sequences for S p-fixable in the ADMG $\mathcal{G}(V)$, then for any $p(V \cup H)$ Markov relative to a DAG $\mathcal{G}(V \cup H)$ that yields the latent projection $\mathcal{G}(V)$, $\Phi_{\sigma_S^1}(p(V); \mathcal{G}(V)) = \Phi_{\sigma_S^2}(p(V); \mathcal{G}(V)) = p(V \setminus S \mid \text{do}(S = s))$, if $\mathcal{G}(V \cup H)$ is interpreted as a causal diagram.



(a) $\mathcal{G}$ (b) $\Phi_{Z_2}(\mathcal{G})$ (c) $\Phi_{\{Z_1,Z_2\}}(\mathcal{G})$ (d) $\Phi_{Z_1}(\mathcal{G})$

Figure 13: (a) An ADMG where T is not fixable. (b, c) A valid sequence of fixing that yields a CADMG where T is p-fixable and $p(Y(t))$ can be obtained as $p(Y(t)) = p(Y(t, z_1, z_2))$. (d) A graph obtained after Z_1 is p-fixed, where $p(Y(t))$ can be obtained by p-fixing T, as $p(Y(t)) = p(Y(t, z_1))$.

Due to this lemma, for marginals $p(V)$ induced by causal models associated with a hidden variable DAG $\mathcal{G}(V \cup H)$, if S is p-fixable in the latent projection $\mathcal{G}(V)$, we can define $\Phi_S(p(V); \mathcal{G}(V))$ to be the result of applying the fixing operator Φ to any p-fixable sequence on S and $p(V), \mathcal{G}(V)$. As suggested by the result, the proof of commutativity comes from viewing each valid p-fixing operation as a step in an identification procedure for the post intervention distribution $p(V \setminus S \mid \text{do}(S = s))$. Thus, any valid sequence for S should yield the same post intervention distribution. This also suggests the following general procedure for identification of the target parameter via a sequence of p-fixing operations.

Corollary 16 (Identification via a sequence of p-fixing)
Fix a causal model associated with a hidden variable DAG $\mathcal{G}(V \cup H)$, that induces the observed marginal distribution $p(V)$. Given $Y^* \equiv \text{an}_{\mathcal{G}_{V \setminus T}}(Y)$, $\psi(t)$ is identified if there exists a subset $Z \subseteq V \setminus (Y^* \cup T)$ that is p-fixable in $\mathcal{G}(V)$ such that T is p-fixable in $\Phi_Z(\mathcal{G}(V))$. When $\psi(t)$ is identified in this manner we have,

$$\psi(t) = \sum_{V \setminus \{Z \cup T\}} Y \times \Phi_{Z \cup T}(p(V); \mathcal{G}(V))\Big|_{T=t}. \tag{39}$$

The result in Corollary 16 directly yields an inverse weighted estimator. We illustrate this via the following example.

Example: Identification via a sequence of p-fixing steps

Consider the ADMG shown in Fig. 13(a). Clearly T is not p-fixable as it does not meet the condition that $\text{dis}_{\mathcal{G}}(T) \cap \text{ch}_{\mathcal{G}}(T) = \emptyset$. However, Z_1 is p-fixable, and yields a CADMG $\Phi_{Z_1}(\mathcal{G})$ where T is p-fixable as shown in Fig. 13(d). Further, Z_1 is an ancestor of Y, but only via a directed path through T. Hence, while the CADMG in Fig. 13(d) corresponds to the post-intervention distribution $p(C, Z_2(z_1), T(z_1), Y(z_1))$, fixing T in this CADMG yields the

post-intervention distribution $p(C, Z_2(z_1), Y(t))$ from which $p(Y(t))$ can be easily obtained as $Y(t, z_1) = Y(t)$ (Malinsky et al., 2019). A similar argument can be made to show that p-fixing according to the sequence (Z_2, Z_1), resulting in the ADMGs shown in Figs. 13(b, c), also gives us the desired post-intervention distribution as $p(C, Y(t, z_1, z_2)) = p(C, Y(t))$.

Consider the first scenario where we p-fix Z_1 prior to p-fixing T in the graph $\Phi_{Z_1}(\mathcal{G}(V))$ shown in Fig. 13(d). We can estimate $\psi(t)$ via the following estimating equation:

$$\mathbb{P}_n\left[p^*(Z_1) \times \frac{\sum_{Z_1} p(Y \mid T, Z, C) \times p(T \mid Z, C) \times p(Z_1)}{p(Y \mid T, Z, C) \times p(T \mid Z, C) \times p(Z_1)} \times U\big(\psi(t); \Phi_{Z_1}(p(V); \mathcal{G}(V))\big)\right] = 0,$$

where $U(\psi(t); \Phi_{Z_1}(p(V); \mathcal{G}(V)))$ is given in Theorem 8, with nuisance models fitted with respect to the weighted distribution $p(V)/\pi_{Z_1}$, rather than with respect to $p(V)$. In this example, the nuisances are simply the propensity score model for the treatment given covariates C and the outcome regression model for Y given the treatment and C, as seen from the CADMG in Fig. 13(d). Further, π_{Z_1} is defined as follows ($Z = \{Z_1, Z_2\}$).

$$\pi_{Z_1} = \frac{p(Y \mid T, Z, C) \times p(T \mid Z, C) \times p(Z_1)}{\sum_{Z_1} p(Y \mid T, Z, C) \times p(T \mid Z, C) \times p(Z_1)}.$$

An estimation strategy is similar to the one used in marginal structural models (Robins, 2000). That is, we can fit a weighted regression for $\mathbb{E}[Y \mid T = t, C]$ using weights $1/\pi_{Z_1}$ estimated via appropriate nuisance models, a similarly weighted model for $p(A \mid C)$, and then plugging these in to solve the final estimating equation.

Using the alternative p-fixing sequence (Z_2, Z_1) also yields a CADMG $\Phi_{Z_1,Z_2}(\mathcal{G})$ where T is p-fixable as in Fig. 13(c). In this case, π_{Z_2} is simply $p(Z_2 \mid Z_1)$. In the CADMG $\Phi_{Z_2}(\mathcal{G})$, the variable Z_1 is childless. Therefore, p-fixing Z_1 in the corresponding distribution corresponds to marginalization of Z_1 (Richardson et al., 2017). Any p-fixings that correspond to marginalization in this manner, do not require the specification of an additional p-fixing weight. Hence, the estimating equation in this case is simply,

$$\mathbb{E}\left[\frac{p^*(Z_1, Z_2)}{p(Z_2 \mid Z_1)} \times U\big(\psi(t); \Phi_{Z_1,Z_2}(p(V); \mathcal{G}(V))\big)\right] = 0. \tag{40}$$

A similar strategy for estimation can be followed here using the weights $1/\pi_{Z_2}$.

In cases where multiple p-fixing operations must be performed in some sequence, say $(Z_1, \ldots, Z_k)$, the first set of weights $1/\pi_{Z_1}$ are obtained by fitting appropriate nuisance models using the observed data. Subsequent weights $1/\pi_{Z_i}$ for $1 < i \leq k$ are obtained by fitting weighted regressions that use a product of the prior weights $1/(\pi_{Z_1} \times \cdots \times \pi_{Z_{i-1}})$. The final nuisance models in $U(\psi(t); \Phi_Z(p(V); \mathcal{G}(V)))$ are then fit using weights $1/(\pi_{Z_1} \times \cdots \times \pi_{Z_k})$. The outer expectation is simply evaluated empirically.

An interesting special case is when the nuisance models for π_Z can be estimated using pieces of the observed data likelihood that are variationally independent from the final nuisance models in $U(\cdot; \cdot)$. The resulting estimators in these cases exhibit double robustness *after* correct specification of models involved in estimating π_Z. In the above example, when we use the order (Z_2, Z_1) we only had to estimate weights $1/\pi_{Z_2}$. Since Z_2 is in a different district from Y and T, we obtain the desired variational independence for any natural parameterization of the observed data likelihood. However, generally (e.g., when we first p-fix Z_1 above) the exact form of robustness is an interesting question for future work.

G. Details on Simulated Data

The R code is available upon request.

To generate data from the ADMGs in Figures 2(b) and 3, we first generated six hidden variables that are used across the first three simulations: H_1 and H_4 are sampled from a Binomial distribution, with $p(H_1 = 1) = 0.4$ and $p(H_4 = 1) = 0.6$. H_2 and H_5 are sampled from a Uniform distribution with corresponding lower bounds $0, -1$ and upper bounds $1.5, 1$. H_3 and H_6 are sampled from a Normal distribution with means 0 and standard deviations 1 and 1.5. For the observed variables in each simulation, continuous variables are sampled from either a Normal or Uniform distributions and binary variables are sampled from Bernoulli distributions. In both ADMGs, we assume all variables, except for the outcome and part of the baselines, are binary random variables. In Fig. 2(b), we assume we have three baselines C_1, C_2, C_3 with Binomial ($p = 0.3$), Uniform ($l = -1, u = 2$), and Normal ($\mu = 1, \text{sd} = 1$) distributions. In Fig. 3, we assume C_1 consists of C_{11} and C_{12} with Normal ($\mu = 1, \text{sd} = 1$) and Uniform ($l = -1, u = 1$) distributions, and C_2 consists of C_{21} and C_{22} with standard Normal and Binomial ($p = 0.4$) distributions. We projected out Z_1 and Z_2 from the DGP. Below we illustrate the data generating processes. $F_V(v)$ denotes the CDF of a standard normal distribution.

$$C_4 = F_{C_3}(c_3),\ C_5 = C_3^{C_1} + (1 - C_1) \times \sin(|C_3|\pi),\ C_6 = C_1 \times C_2 + |C_3| \qquad \text{(Fig. 2(b))}$$
$$p(T = 1 \mid C, U) \sim \text{expit}(0.5 + 0.9C_4 - 0.5C_5 + 0.2C_6 + 0.3U_1 - 0.8U_2 + 0.8U_3)$$
$$p(M = 1 \mid C, T, U) \sim \text{expit}(0.5 - 0.7C_1 + 0.8C_2 - C_3 - 1.2T - 0.2U_4 + 0.5U_5 + 0.4U_6 + (1.5C_4 + 1.2C_5 + 0.6C_6)T)$$
$$p(L = 1 \mid C, M, U) \sim \text{expit}(-0.5 + 0.8C_4 + 1.2C_5 - 0.6C_6 - 1.2M + 0.3U_1 + 0.6U_2 - 0.4U_3 - (0.8C_4 + 1.5C_5 + 0.4C_6)M)$$
$$Y \mid C, T, L, U \sim 0.5 + 0.5C_4 - 2C_5 + 0.8C_6 + 0.5T + 0.6L - 0.6U_4 + 0.5U_5 - 0.5U_6 + 1.3C_4A + 2.3C_5L + 2C_6TL + 1.2AL + \mathbb{N}(0, 1.5).$$

$$C_3 = F_{C_{11}}(c_{11}c_{12}) + (1 - C_{12}) \times \sin(|C_{11}|\pi),\ C_4 = (C_{21}^{C_{22}}) + (1 - C_{22}) \times \sin(|C_{21}|\pi)) \qquad \text{(Fig. 3)}$$
$$p(T = 1 \mid C, U) \sim \text{expit}(-0.5 + 0.9C_{11} - 0.7C_{12} + 0.6C_{21} - 0.7C_{22} + 0.3U_1 - 0.5U_2 + 0.4U_3 + 1.6C_3 - 0.8C_4)$$
$$p(M = 1 \mid C, T) \sim \text{expit}(-0.5 - 1.4C_{21} + 1.3C_{22} - 1.2A + 2.2C_4A - C_4)$$
$$p(L = 1 \mid C, M, U) \sim \text{expit}(0.5 - 0.5 * C_{11} - 0.4C_{12} + 0.8C_{21} + 0.9C_{22} - 1.2M + 0.3U_1 + 0.6U_2 - 0.4U_3 - 1.8C_3M - 1.5C_4M + 1.2C_3 + 0.8C_4)$$
$$Y \mid C, L \sim 0.5 + 0.7C_{21} - 0.5C_{22} + 1.6L + 1.1C_4L + 0.8C_4 + \mathbb{N}(0, 1.5).$$

To generate data from the ADMG in Figure 5, we first generated ten hidden variables that are used for the last simulation: H_1, H_3, H_5, H_7, H_9 are sampled from a Binomial distribution with $p_1 = 0.4, p_3 = 0.3, p_5 = 0.4, p_7 = 0.3, p_9 = 0.3$. $H_2, H_4, H_6, H_8, H_{10}$ are sampled from standard Normal distribution. We assume C consists of two baseline covariates. All variables, except for the baselines and outcome, are binary random variables.

$$p(R_2 = 1 \mid U) \sim \text{expit}(-0.2 + 0.3U_1 - 0.8U_2 + 0.4U_3 + 0.6U_4) \qquad \text{(Fig. 5)}$$
$$C_1 \mid U \sim 0.4U_7 - 0.1U_8 + 0.6U_9 + 0.8U_{10} + \mathbb{N}(0, 1)$$
$$C_2 \mid U \sim -0.3U_7 - 0.7U_8 + 0.8U_9 + 1.2U_{10} + \mathbb{N}(0, 1)$$
$$C_3 = |C_1C_2|^{0.5} + \sin(|C_1 + C_2|\pi),\ C_4 = F_{C_1}(c_1),$$
$$p(Z = 1 \mid U) \sim \text{expit}(-0.5 + U_1 + 0.2U_2 - 0.8U_5 + 0.3U_6)$$
$$p(T = 1 \mid C, Z, U) \sim \text{expit}(0.5 - 0.5C_1 + 0.5C_2 + 0.3Z + 0.5U_3 - 0.4U_4 + 0.8C_3 - 1.3C_4)$$
$$p(R_1 = 1 \mid T, U) \sim \text{expit}(0.2 + 0.7T - 0.6U_5 - 0.6U_6)$$
$$p(M = 1 \mid R_1, U) \sim \text{expit}(0.5 - 0.8R_1 + 1.2U_7 - 1.5U_8)$$
$$Y \mid R_2, C, T, M, U \sim -1 + 0.5C_1 + 0.2C_2 + 1.2T + 0.8R_2 + 0.8M + 0.2U_9 - 0.4U_{10} + 0.8C_3 - 1.2C_4 + MA + \mathbb{N}(0, 1).$$

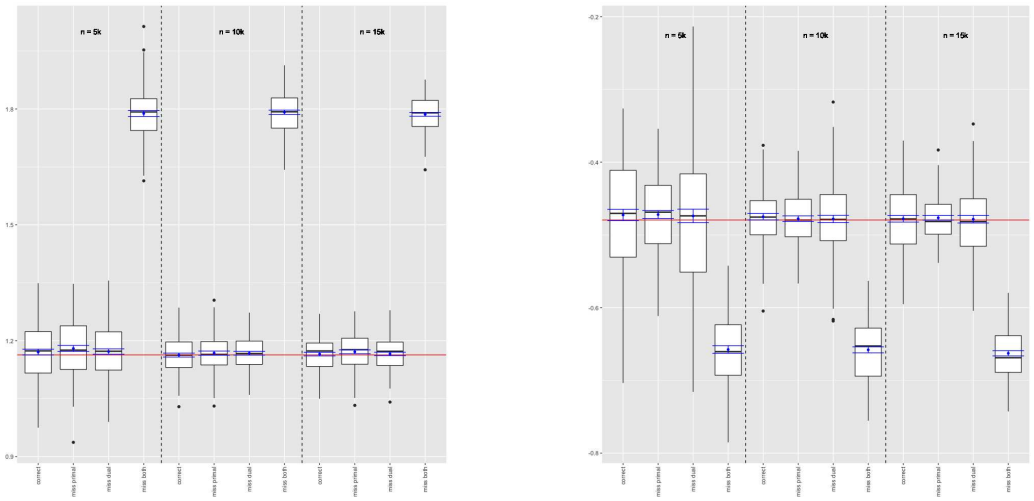


Figure 14: Double robustness of the Augmented Primal IPW estimator. The boxplot panel on the (left) uses the ADMG in Fig. 2(b), and the one on the (right) uses the ADMG in Fig. 3. The red dashed lines indicate the true values of the ACE. The blue dots denote the mean and the blue error bars denote the length of the standard error of the mean.

Additional Experiments: In continuation of Simulation 2, in Figure 14 we highlight the summary statistics for the case of using Augmented Primal IPW estimator for computing the causal effect of T on Y in ADMGs of Figures 2(b) and 3.

We reran the simulations for primal fixability with various different DGPs (different parameter sets.) Our conclusions remain consistent across all simulations. It is worth pointing out that in general, IPW-type estimators are instable due to extreme values for the weights used in the estimation procedures. The followings are examples of two DGPs (using the ADMGs in Figures 2(b) and 3, where we encountered some instability in the weights and consequently in how primal and dual IPW estimators behave as sample size increases. We plot the bias against the sample size in Figure 15. We truncated the extreme weights in both examples and were able to observe a more smoothed behavior, however, such procedures introduce bias that does not disappear as sample size increases. This is shown in Figure 16. Exploring possibilities for dealing with such instabilities is an interesting future direction.

$$C_4 = F_{C_3}(c_3),\ C_5 = C_3^{C_1} + (1 - C_1) \times \sin(|C_3|\pi),\ C_6 = C_1 \times C_2 + |C_3| \qquad \text{(Fig. 2(b))}$$
$$p(T = 1 \mid C, U) \sim \text{expit}(0.5 + 0.9C_4 - 0.5C_5 + 0.2C_6 + 0.3U_1 - 0.8U_2 + 0.8U_3)$$
$$p(M = 1 \mid C, T, U) \sim \text{expit}(0.5 - 0.7C_1 + 0.8C_2 - C_3 - 1.2T - 0.2U_4 + 0.5U_5 + 0.4U_6 + (1.5C_4 + 1.2C_5 + 1.6C_6)T)$$
$$p(L = 1 \mid C, M, U) \sim \text{expit}(-0.5 + 0.8C_4 + 1.2C_5 - 0.6C_6 - 1.2M + 0.3U_1 + 0.6U_2 - 0.4U_3 - (0.8C_4 + 1.5C_5 + 0.4C_6)M)$$
$$Y \mid C, T, L, U \sim 0.5 + 0.5C_4 - 2C_5 + 0.8C_6 + 0.5T + 0.6L - 0.6U_4 + 0.5U_5 - 0.5U_6 + 1.3C_4A + 2.3C_5L + 2C_6TL - 1.2AL + \mathbb{N}(0, 1.5).$$
$$C_3 = F_{C_{11}}(c_{11}),\ C_4 = (C_{21}^{C_{22}}) + (1 - C_{21}) \times \sin(|C_{21}|\pi)) \qquad \text{(Fig. 3)}$$
$$p(T = 1 \mid C, U) \sim \text{expit}(0.5 + 0.9C_{11} - 0.5C_{12} + 0.2C_{21} + 1.2C_{22} + 0.3U_1 - 0.8U_2 + 0.8U_3 + 0.7C_3 - 2C_4)$$
$$p(M = 1 \mid C, T) \sim \text{expit}(0.5 - 0.7C_3 + 0.8C_4 - 1.2A + 1.5C_3A + 1.2C_4A)$$
$$p(L = 1 \mid C, M, U) \sim \text{expit}(-0.5 - 0.5C_3 - 0.1C_{12} + 0.8C_{21} + 1.2C_{22} + 0.3U_1 + 0.6U_2 - 0.4U_3 - (1.2 + 0.8C_{21} + 1.5C_{22})M)$$
$$Y \mid C, L \sim 0.5 + 0.5C_{21} - 2C_{22} + 0.6L + 1.3C_{21}L + 2.3C_{22}L + \mathbb{N}(0, 1.5).$$

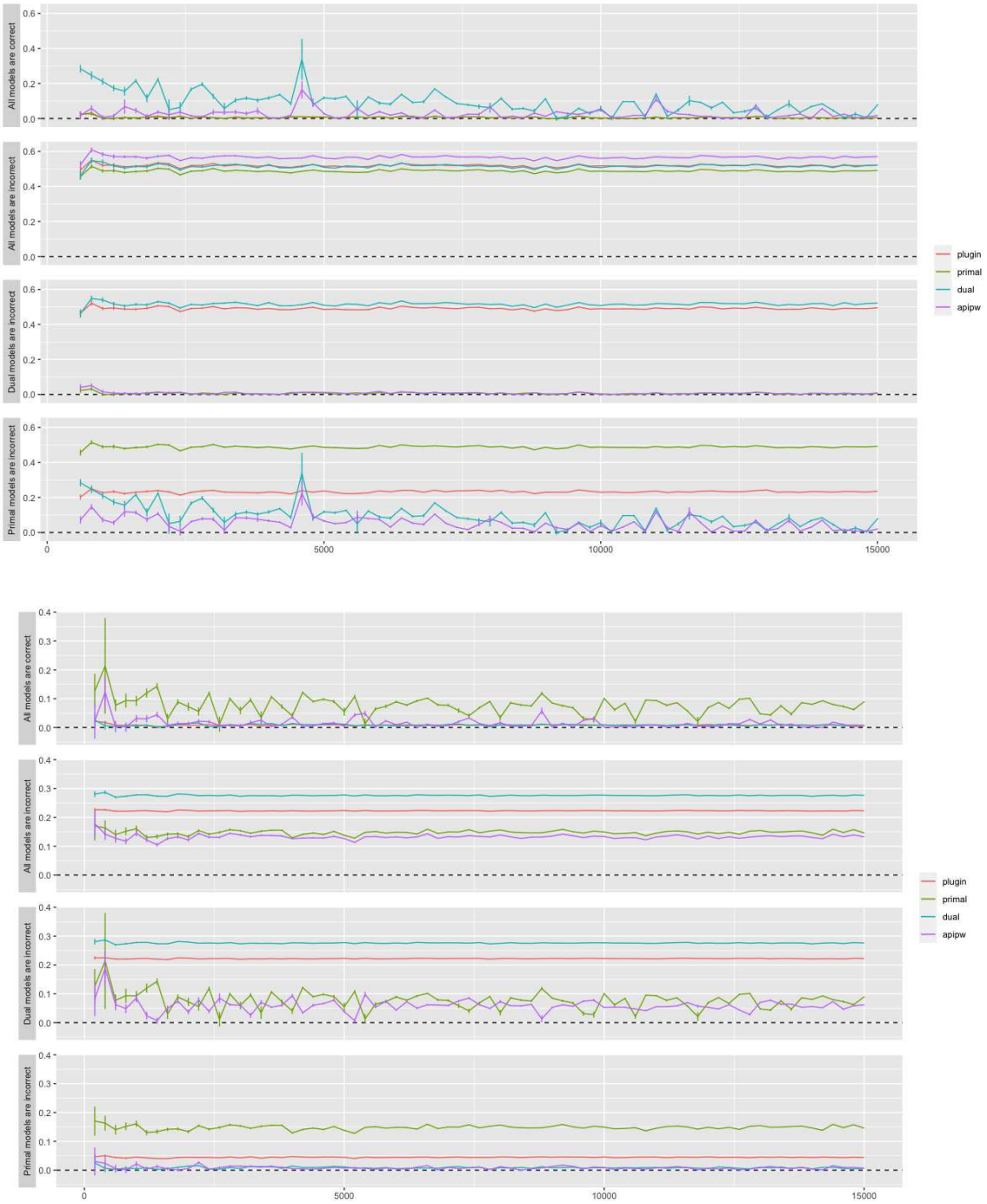


Figure 15: Bias behavior as a function of sample size with an alternative set of parameters for the DGPs that can result in extreme weights in the IPW estimators. (top) using the ADMG in Fig. 2(b). (bottom) using the ADMG in Fig. 3.

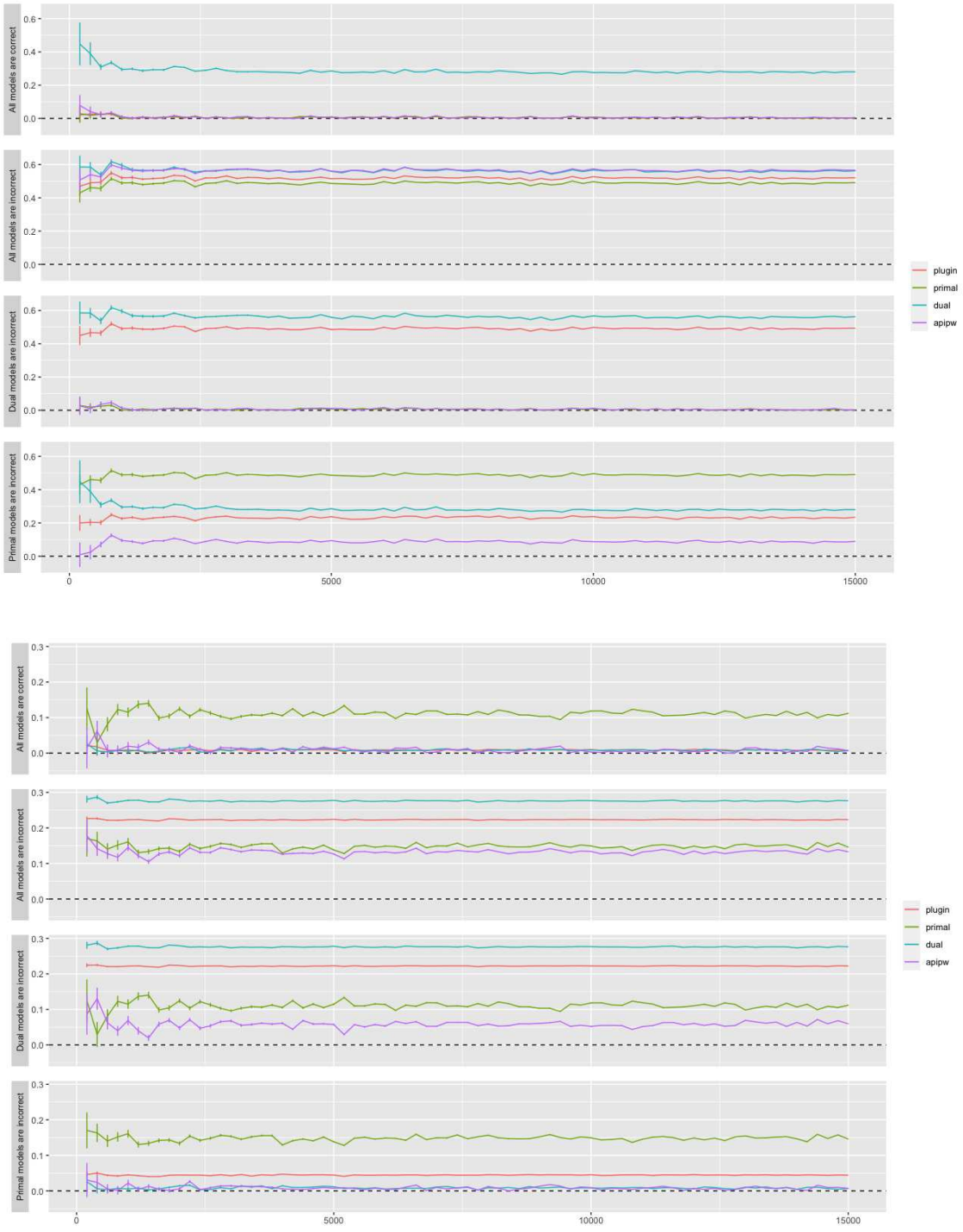


Figure 16: Bias behavior as a function of sample size with truncated weights in Fig 15. (top) using the ADMG in Fig. 2(b). (bottom) using the ADMG in Fig. 3. The plots have fewer outliers. However, truncating weights in primal and dual terms introduces bias that does not disappear as sample size increases.

H. Proofs

Theorem 1 (Soundness and completeness of Algorithm 1)

Proof

The construction of Algorithm 1 is closely related to the maximal arid projection described by Shpitser et al. (2018). Given any ADMG $\mathcal{G}$ the maximal arid projection yields a maximal arid graph (MArG) $\mathcal{G}^a$ that is nested Markov equivalent to $\mathcal{G}$. The MArG $\mathcal{G}^a$ has

1. A directed edge $V_j \to V_i$ if and only if $V_j \in \text{an}_{\mathcal{G}}(V_i)$ and V_j is a parent of any element in the *reachable closure* of V_i – the set of vertices that are still random in $\phi_{\neg\{V_i\}}(\mathcal{G})$.

2. A bidirected edge $V_i \leftrightarrow V_j$ if and only if the previous condition fails, and $\phi_{\neg\{V_i,V_j\}}(\mathcal{G})$ contains a single district.

We now prove that Algorithm 1 declares $\mathcal{G}$ to be NPS if and only if $\mathcal{G}^a$ is a complete graph (no missing edges). We will then use this observation to reason about the soundness and completeness of the algorithm. Notice that for any pair (V_i, V_j) and assuming wlog that $V_j \prec_\tau V_i$, the first check in line 4 of the algorithm evaluates to True when V_j is not a parent of the reachable closure of V_i, as the district of V_i in $\phi_{\neg\{V_i\}}(\mathcal{G})$ is the same as the reachable closure of V_i (elements of the district of V_i are the only vertices that can remain random in such a CADMG.) That is, this condition evaluates to True when $V_j \to V_i$ is not added to $\mathcal{G}^a$ fails. It is also easy to see that the second check in line 4 evaluates to True when $V_i \leftrightarrow V_j$ is not added to $\mathcal{G}^a$. Hence, when both checks evaluate to True the algorithm returns "not NPS" and the MArG $\mathcal{G}^a$ has at least one pair (V_i, V_j) that are not adjacent. On the other hand, when the truth value output in line 4 evaluates to False for every pair (V_i, V_j) resulting in the algorithm returning "NPS", it is clear that either V_j is a parent of the reachable closure of V_i, or if not, $\phi_{\neg\{V_i,V_j\}}(\mathcal{G})$ contains a single district. That is, $\mathcal{G}^a$ either contains $V_j \to V_i$ or $V_j \leftrightarrow V_i$ for every pair when the algorithm returns "NPS."

Completeness

We now prove that the algorithm is complete – when the algorithm declares the model of $\mathcal{G}(V)$ to be NPS it is indeed NPS. We have seen that when the algorithm returns NPS, the input ADMG $\mathcal{G}$ is nested Markov equivalent to a MArG $\mathcal{G}^a$ where every pair of vertices is pairwise connected with a directed or bidirected edge. The nested Markov model of any complete ADMG is NPS (per Corollary 17), and so it is indeed the case that $\mathcal{G}$ is NPS, and the output of the algorithm is correct.

Soundness

The algorithm returns "not NPS" when the input ADMG $\mathcal{G}$ is nested Markov equivalent to a MArG $\mathcal{G}^a$ where at least one pair of vertices, say (V_i, V_j), are not connected by a directed or bidirected edge. Assume wlog that $V_i \prec V_j$ according to a valid topological order, and (V_i, V_j) is the only non-adjacent pair in $\mathcal{G}^a$. We assume the edge between V_i and V_j is the only missing edge for simplicity, but it is easy to see if there are additional missing edges this simply corresponds to a submodel where there is still a constraint between V_i and V_j.

We know the complete graph where the edge is present between V_i and V_j is NPS (no constraints in the nested Markov model.) We now ask whether the constraint finding

algorithm in Tian and Pearl (2002b), which returns a list of constraints that imply the nested Markov model of an ADMG, detects a constraint between V_i and V_j in $\mathcal{G}$, implying it is not NPS.

The algorithm in Tian and Pearl (2002b), and reformulated in terms of kernels and fixing in Richardson et al. (2017), starts by examining the subgraph of $\mathcal{G}$ consisting of V_j and all vertices preceding V_j under a valid topological order, namely $\mathcal{G}_{\{\preceq_\tau V_j\}}$. Let S be the district of $\mathcal{G}_{\{\preceq_\tau V_j\}}$ containing V_j. If V_i is not in the district S nor a parent of the district S, then we immediately get an ordinary conditional independence constraint $V_i \perp\!\!\!\perp V_j \mid \mathrm{mp}_{\mathcal{G}}(V_j)$ in step (A1) of the algorithm. We now examine the recursion that occurs in step (A2) to handle the remaining cases: (i) V_i is in the district S, but not a parent of S, and (ii) V_i is a parent of S and potentially in the district S. Recursive applications of step (A2) in the algorithm involve examining CADMGs with random vertices that are subsets of S, and corresponding kernels obtained from $p(V)$, obtained by one of two steps. Given a recursively obtained CADMG $\mathcal{G}(\tilde{S}, W)$, the first step considers subgraphs $\mathcal{G}_A$ of $\mathcal{G}(\tilde{S}, W)$ consisting of all ancestral subsets A of $\tilde{S} \cup W$ in $\mathcal{G}(\tilde{S}, W)$ that contains V_j.[10] The second step considers a district E containing V_j within $\mathcal{G}_A$. These steps, occuring within recursive applications of (A2), visit every intrinsic set in the original ADMG $\mathcal{G}(V)$, potentially multiple times.

Case (i): V_i and V_j belong to the same district in $\mathcal{G}_{\{\preceq_\tau V_j\}}$. Consider a sequence of alternating ancestral and district steps within (A2), starting from S, which always include $\{V_i, V_j\}$ in the random vertex set $\tilde{S}$ for the CADMG $\mathcal{G}(\tilde{S}, \tilde{W})$ associated with each step. This sequence is non-empty, since V_i and V_j are in S, by assumption. Moreover, V_i and V_j will always be childless in each CADMG $\mathcal{G}(\tilde{S}, \tilde{W})$. Since the projection $\mathcal{G}^a$ is a MArG where V_i and V_j are not adjacent, we know $\phi_{\neg\{V_i,V_j\}}(\mathcal{G}_S) = \phi_{\neg\{V_i,V_j\}}(\mathcal{G})$ is a CADMG where V_i and V_j are not bidirected connected. Thus, there must exist a point in this sequence where V_i and V_j will no longer be in the same district after an ancestral step applied to some $\mathcal{G}(\tilde{S}, \tilde{W})$, where an ancestral set A is retained. Since V_i is childless in the resulting CADMG, V_i is not in the Markov blanket of V_j, and the algorithm in Tian and Pearl (2002b) adds the corresponding constraint, namely that the kernel $q(\tilde{E} \mid \mathrm{pa}_{\mathcal{G}}(\tilde{E}))$ is not a function of V_i, where $\tilde{E}$ is the district containing V_j in $\mathcal{G}(\tilde{S}, \tilde{W})_A$, and $q(\tilde{E} \mid \mathrm{pa}_{\mathcal{G}}(\tilde{E}))$ is the corresponding kernel obtained from $\sum_{\tilde{S}\setminus A} q(\tilde{S} \mid \mathrm{pa}_{\mathcal{G}}(\tilde{S}))$.

Case (ii): V_i is a parent of the district S and potentially also a part of the district S in $\mathcal{G}_{\{\preceq_\tau V_j\}}$. Consider a sequence of alternating ancestral and district steps within (A2), starting from S, which always include $V_j \in \tilde{S}$ and $V_i \in \tilde{S} \cup \tilde{W}$ for the CADMG $\mathcal{G}(\tilde{S}, \tilde{W})$ associated with each step. This sequence is non-empty, since V_i is in $\mathrm{pa}_{\mathcal{G}}(S)$. Since the projection $\mathcal{G}^a$ is a MArG where V_i and V_j are not adjacent, we know $\phi_{\neg\{V_j\}}(\mathcal{G})$ is a CADMG where V_i is not a parent of the district of V_j. Thus, there must exist a point in this sequence where after either an ancestral or district step, V_i will no longer be a parent of the district of V_j in the CADMG $\mathcal{G}(\tilde{S}, \tilde{W})_A$ (if an ancestral step was taken) or $\mathcal{G}(\tilde{E}, \tilde{W})_{\tilde{E}}$ (if the district step was taken). At this point, the algorithm in Tian and Pearl (2002b) adds the corresponding constraint, namely that the kernel $\sum_{\tilde{S}\setminus A} q(\tilde{S} \mid \mathrm{pa}_{\mathcal{G}}(\tilde{S}))$ (if an ancestral step was taken) or $q(\tilde{E} \mid \mathrm{pa}_{\mathcal{G}}(\tilde{E}))$ (if a district step was taken) is not a function of V_i.

10. An ancestral margin of an CADMG $\mathcal{G}$ is a subgraph with a vertices that is closed under the ancestral relation in $\mathcal{G}$.

Hence, we see that in either case, the algorithm gives a non-empty list of constraints associated with $\mathcal{G}$ whenever there is an edge missing between V_i and V_j in the MArG projection $\mathcal{G}^a$. ■

Theorem 2 (mb-shielded ADMGs)

Proof The proof relies on the fact that the constraint finding algorithm provided in Tian and Pearl (2002b) finds a list of equality constraints that is sufficient to define the nested Markov model of an ADMG (this was shown by Richardson et al. (2017)). Here we show that the only non-trivial equality constraints found by applying this algorithm to an arbitrary mb-shielded ADMG $\mathcal{G}$ are of the form $V_i \perp\!\!\!\perp \{\prec V_i\} \mid \mathrm{mp}_{\mathcal{G}}(V_i)$, for a topological order $\prec$, thus implying that all equality constraints in the nested Markov model of such an ADMG are implied by ordinary conditional independences of that form.

Given a topological order $\prec$ on the vertices in an ADMG $\mathcal{G}(V)$, the constraint finding algorithm in Tian and Pearl (2002b) iterates over each vertex V_i in the order and attempts to find constraints between V_i and $\{\prec V_i\}$. In substep (A1) of the algorithm (see Tian and Pearl (2002b) for details), it identifies constraints of the form $V_i \perp\!\!\!\perp \{\prec V_i\} \mid \mathrm{mp}_{\mathcal{G}}(V_i)$. Step (A2) and recursive applications of it, attempts to find constraints between V_i and subsets of $\{\prec V_i\} \cap \mathrm{mb}_{\mathcal{G}}(V_i)$. In the rest of this proof, we temporarily switch to using a disjunctive definition of parents, i.e., $\mathrm{pa}_{\mathcal{G}}(S) = \bigcup_{S_i \in S} \mathrm{pa}_{\mathcal{G}}(S_i)$, to align with the presentation of the constraint finding algorithm in Tian and Pearl (2002b).

The top level calls to step (A2) always involve a subset S of a district D of $\mathcal{G}(V)$, along with $\mathrm{pa}_{\mathcal{G}(V)}(S) \setminus S$. These subsets S are districts of V_i in subgraphs $\mathcal{G}_{\{\prec V_i\}}$ of $\mathcal{G}(V)$. Each call to (A2) involves constraints on a particular V_i. Since $\mathcal{G}(V)$ is mb-shielded, the induced subgraph $\mathcal{G}(V)_{S \cup \mathrm{pa}_{\mathcal{G}(V)}(S)}$ is complete. In particular, since every $V_j \in \mathrm{pa}_{\mathcal{G}(V)}(S) \setminus S$ is in the Markov blanket of every $V_k \in S$, $V_j \in \mathrm{pa}_{\mathcal{G}(V)}(V_k)$. The top level calls to step (A2) proceed as follows.

First, a subset $\tilde{S}$ of S ancestral in $\mathcal{G}(V)_{S \cup \mathrm{pa}_{\mathcal{G}(V)}}(S)$ is found. A constraint is found if $\{[\mathrm{pa}_{\mathcal{G}(V)}(S) \cup S] \setminus (S \setminus \tilde{S})\} \setminus [\mathrm{pa}_{\mathcal{G}(V)}(\tilde{S}) \cup \tilde{S}] \neq \emptyset$. Note that $\{[\mathrm{pa}_{\mathcal{G}(V)}(S) \cup S] \setminus (S \setminus \tilde{S})\} = ([\mathrm{pa}_{\mathcal{G}(V)}(S) \cup S] \cap \tilde{S}) \cup (\mathrm{pa}_{\mathcal{G}(V)}(S) \setminus S)$. Since $\mathcal{G}(V)$ is mb-shielded, and every element $V_j \in \mathrm{pa}_{\mathcal{G}(V)}(S) \setminus S$ is in the Markov blanket of every element in S (and thus in $\tilde{S}$) in $\mathcal{G}_{\{V_i\}}$. Thus, every element $V_j \in \mathrm{pa}_{\mathcal{G}(V)}(S) \setminus S$ is in $\mathrm{pa}_{\mathcal{G}(V)}(V_k)$ for every $V_k \in \tilde{S}$. Any element V_j in $[\mathrm{pa}_{\mathcal{G}(V)}(S) \cup S] \cap \tilde{S}$ must be in $\mathrm{pa}_{\mathcal{G}(V)}(\tilde{S}) \cup \tilde{S}$ by definition. Consequently, $\{[\mathrm{pa}_{\mathcal{G}(V)}(S) \cup S] \setminus (S \setminus \tilde{S})\} \setminus [\mathrm{pa}_{\mathcal{G}(V)}(\tilde{S}) \cup \tilde{S}] = \emptyset$, and no restrictions are added at this stage.

Second, the algorithm considers a district $\tilde{E}$ of V_i in $\mathcal{G}_{\tilde{S} \cup \mathrm{pa}_{\mathcal{G}(V)}(\tilde{S})}$, and $\mathrm{pa}_{\mathcal{G}(V)}(\tilde{E})$. A constraint is found if $[\mathrm{pa}_{\mathcal{G}(V)}(\tilde{S}) \cup \tilde{S}] \setminus [\mathrm{pa}_{\mathcal{G}(V)}(\tilde{E}) \cup \tilde{E}] \neq \emptyset$. Any element V_j in $\mathrm{pa}_{\mathcal{G}(V)}(\tilde{S}) \setminus \tilde{S}$ must be in $\mathrm{pa}_{\mathcal{G}}(S) \setminus S$, since $\tilde{S}$ is ancestral in $\mathcal{G}(V)_{S \cup \mathrm{pa}_{\mathcal{G}(V)}}(S)$. Consequently $V_j \in \mathrm{pa}_{\mathcal{G}(V)}(V_k)$ for every $V_k \in \tilde{E} \subseteq S$.

Any element $V_j \in \tilde{S}$ is in the Markov blanket of every element in $\tilde{E}$, in particular V_i. Thus, V_j and V_i must share an edge. Since V_i is the $\prec$-largest element in S under a topological order, this edge must be $V_j \to V_i$ or $V_j \leftrightarrow V_i$. In the former case, $V_j \in \mathrm{pa}_{\mathcal{G}(V)}(\tilde{E})$,

while the latter case, $V_j \in \tilde{E}$. In either case, $[\mathrm{pa}_{\mathcal{G}(V)}(\tilde{S}) \cup \tilde{S}] \setminus [\mathrm{pa}_{\mathcal{G}(V)}(\tilde{E}) \cup \tilde{E}] = \emptyset$, and no restrictions are added at this stage.

Step (A2) is then recursively applied to $\mathcal{G}(V)_{\tilde{E} \cup \mathrm{pa}_{\mathcal{G}(V)}(\tilde{E})}$. Since for any $V_j, V_k \in \tilde{E} \cup \mathrm{pa}_{\mathcal{G}(V)}(\tilde{E})$, one of the pair is in the Markov blanket of the other, they must share an edge in $\mathcal{G}(V)_{\tilde{E} \cup \mathrm{pa}_{\mathcal{G}(V)}(\tilde{E})}$. Moreover, any element in $\mathrm{pa}_{\mathcal{G}(V)}(\tilde{E}) \setminus \tilde{E}$ must be connected to any elements in $\tilde{E}$ by a directed edge. This allows us to repeat the above argument inductively, with $\tilde{E}$ replacing S and $\mathrm{pa}_{\mathcal{G}(V)}(\tilde{E}) \setminus \tilde{E}$ replacing $\mathrm{pa}_{\mathcal{G}(V)}(S) \setminus S$, to establish that no constraints are added on any recursive call in step (A2).

Thus, running the algorithm on an mb-shielded ADMG returns a list of constraints consisting of only ordinary conditional independence constraints (those that are found by substep (A1)), and specifically ones that are of the form $V_i \perp\!\!\!\perp \{\prec V_i\} \mid \mathrm{mp}_{\mathcal{G}}(V_i)$. ■

Corollary 17 (Complete ADMGs are nonparametric saturated)
Any complete ADMG $\mathcal{G}(V)$ is mb-shielded, and observationally equivalent to a complete DAG. Hence, the nested Markov model of a complete ADMG is nonparametric saturated.

Proof The first claim follows from the definition of mb-shieldedness, and by application of Theorem 2. Since complete DAGs are nonparametric saturated, so is the ADMG $\mathcal{G}(V)$. ■

Lemma 3 (Λ and $\Lambda^{\perp}$ in mb-shielded ADMGs)

Proof The proof here is similar to the proofs of Theorems 4.4 and 4.5 in Tsiatis (2007). Given $p(V)$ that factorizes with respect to an mb-shielded ADMG $\mathcal{G}(V)$, we can write down the following factorization using the ordinary local Markov property $V_i \perp\!\!\!\perp \{\prec V_i\} \setminus \mathrm{mp}_{\mathcal{G}}(V_i) \mid \mathrm{mp}_{\mathcal{G}}(V_i) : p(V) = \prod_{V_i \in V} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))$. Under no restriction, the conditional density $p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)), \forall V_i \in V$, is any positive function such that $\int p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) d\nu(V_i) = 1$, for all values of $\mathrm{mp}_{\mathcal{G}}(V_i))$, where $d\nu(V_i)$ is the dominating measure. Note that $p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))$s are variationally independent.

The tangent space Λ^*, corresponding to the model of the mb-shielded ADMG $\mathcal{G}(V)$, is defined as the mean square closure of all parametric submodel tangent spaces. Assume there are k variables in V. The parametric submodel is defined as $\mathcal{M}_{\mathrm{sub}} = \{\prod_{V_i \in V} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i); \gamma_i)\}$, where $\gamma_i, i = 1, \ldots, k$ are parameters that are variationally independent and $p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i); \gamma_{0i})$ denotes the true conditional density of $p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))$. The parametric submodel tangent space is defined as the space spanned by the joint score $S_\gamma(V_1, \ldots, V_k)$ given as follows,

$$S_\gamma(V_1, \ldots, V_k) = \frac{\partial}{\partial \gamma} \log p(V) = S_{\gamma_1}(V_1) + \cdots + S_{\gamma_k}(V_k, \mathrm{mp}_{\mathcal{G}}(V_k)).$$

Therefore, the parametric submodel tangent space is $\Lambda^*_\gamma = a_1 \times S_{\gamma_1}(V_1) + \cdots + a_k \times S_{\gamma_k}(V_k, \mathrm{mp}_{\mathcal{G}}(V_k))$, where a_i's are constants. Due to variational independence of γ_is, $\Lambda^*_\gamma = \Lambda^*_{\gamma_1} \oplus \cdots \oplus \Lambda^*_{\gamma_k}$, where $\Lambda^*_{\gamma_i} = \{a_i \times S_{\gamma_i}(V_i, \mathrm{mp}_{\mathcal{G}}(V_i))\}$. The tangent space Λ^* is then the mean-square closure of all parametric submodel tangent spaces, i.e., $\Lambda^* = \Lambda^*_1 \oplus \cdots \oplus \Lambda^*_k$, where Λ^*_i is the mean-square closure of the parametric submodel tangent space $\Lambda^*_{\gamma_i}$ which corresponds to the term $p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))$.

By the ordinary local Markov property, i.e., $V_i \perp\!\!\!\perp \{\prec V_i\} \setminus \mathrm{mp}_{\mathcal{G}}(V_i) \mid \mathrm{mp}_{\mathcal{G}}(V_i)$, and properties of score functions for parametric models of conditional densities, the score function $S_i(.)$ must be a function of only $\{V_i, \mathrm{mp}_{\mathcal{G}}(V_i)\}$ and must have conditional expectation $\mathbb{E}[S_i(V_i, \mathrm{mp}_{\mathcal{G}}(V_i)) \mid \mathrm{mp}_{\mathcal{G}}(V_i)] = 0$. Consequently, any element spanned by $S_i(V_i, \mathrm{mp}_{\mathcal{G}}(V_i))$ must belong to Λ_i^*; hence $\Lambda_i^* = \{\alpha(V_i, \mathrm{mp}_{\mathcal{G}}(V_i)) \mid \mathbb{E}[\alpha \mid \mathrm{mp}_{\mathcal{G}}(V_i)] = 0\}$. Further, in order to show that Λ_i^*'s are orthogonal, we need to show that $\mathbb{E}[h_i \times h_j] = 0$, where $h_i \in \Lambda_i^*$ and $h_j \in \Lambda_j^*$,

$$\begin{aligned}
\mathbb{E}\big[h_i \times h_j\big] &= \mathbb{E}\Big[h_i \times \mathbb{E}\big[h_j \mid V_i, \mathrm{mp}_{\mathcal{G}}(V_i)\big]\Big] \\
&= \mathbb{E}\Big[h_i \times \mathbb{E}\Big[\mathbb{E}\big[h_j \mid V_i, \mathrm{mp}_{\mathcal{G}}(V_i), \mathrm{mp}_{\mathcal{G}}(V_j)\big] \;\Big|\; V_i, \mathrm{mp}_{\mathcal{G}}(V_i)\Big]\Big] \\
&= \mathbb{E}\Big[h_i \times \mathbb{E}\Big[\mathbb{E}\big[h_j \mid \mathrm{mp}_{\mathcal{G}}(V_j)\big] \mid V_i, \mathrm{mp}_{\mathcal{G}}(V_i)\Big]\Big] = 0.
\end{aligned}$$

The projection h_i is in Λ_i^*. Therefore, we only need to show that $h - h_i$ is orthogonal to all elements in Λ_i^*. Consider an arbitrary element $\ell \in \Lambda_i^*$,

$$\begin{aligned}
\mathbb{E}[(h - h_i) \times \ell] &= \mathbb{E}\Big[\ell \times \big(\mathbb{E}[h \mid V_i, \mathrm{mp}_{\mathcal{G}}(V_i)] - h_i\big)\Big] = \mathbb{E}[\ell \times \mathbb{E}[h \mid \mathrm{mp}_{\mathcal{G}}(V_i)]] \\
&= \mathbb{E}\Big[\mathbb{E}\big[\ell \times \mathbb{E}[h \mid \mathrm{mp}_{\mathcal{G}}(V_i)] \mid V_i, \mathrm{mp}_{\mathcal{G}}(V_i)\big]\Big] = \mathbb{E}\Big[\mathbb{E}[h \mid \mathrm{mp}_{\mathcal{G}}(V_i)] \times \mathbb{E}\big[\ell \mid V_i, \mathrm{mp}_{\mathcal{G}}(V_i)\big]\Big] = 0.
\end{aligned}$$

Now regarding the orthogonal complement $\Lambda^{*\perp}$ in mb-shielded ADMGs, we know that $\Lambda^{*\perp} = \{h - \pi[h \mid \Lambda^*], \forall h \in \mathbb{H}\}$, by definition. For a given $h \in \mathbb{H}$, we have $h = h_1 + \cdots + h_k$, where $h_i \in \Lambda_i$ as Λ_i is defined in Section 3. According to Section 3, h_i is any function of $V_1, \ldots, V_i$, such that $\mathbb{E}[h_i \mid V_1, \ldots, V_{i-1}] = 0$. Therefore,

$$\begin{aligned}
h - \pi[h \mid \Lambda^*] &= \big(h_1 + \cdots + h_k\big) - \pi[h_1 + \cdots + h_k \mid \Lambda_1^* \oplus \ldots \Lambda_k^*] \\
&= \sum_{i=1}^{k} h_i - \pi[h_i \mid \Lambda_1^* \oplus \ldots \Lambda_k^*] = \sum_{i=1}^{k} h_i - \pi[h_i \mid \Lambda_i^*] \\
&= \sum_{i=1}^{k} h_i - \mathbb{E}[h_i \mid V_i, \mathrm{mp}_{\mathcal{G}}(V_i)] + \mathbb{E}[h_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)].
\end{aligned}$$

The third equality holds since Λ_i is orthogonal to Λ_j^*, for $i, j = 1, \ldots, k$, such that $i \neq j$. Note that $h_i \equiv \mathbb{E}[h(V) \mid V_1, \ldots, V_i] - \mathbb{E}[h(V) \mid V_1, \ldots, V_{i-1}]$. Since $h(V)$ is an arbitrary element of the Hilbert space, without loss of generality, we can replace $\mathbb{E}[h(V) \mid V_1, \ldots, V_i]$ with $\alpha_i(V_1, \ldots, V_i)$. Therefore, $h_i = \alpha_i - \mathbb{E}[\alpha_i \mid V_1, \ldots, V_{i-1}]$. Substituting h_i in the above equation yields the following.

$$\begin{aligned}
&h_i - \mathbb{E}[h_i \mid V_i, \mathrm{mp}_{\mathcal{G}}(V_i)] + \mathbb{E}[h_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)] \\
&\qquad = \alpha_i - \mathbb{E}[\alpha_i \mid V_1, \ldots, V_{i-1}] \\
&\qquad\quad - \mathbb{E}[\alpha_i \mid V_i, \mathrm{mp}_{\mathcal{G}}(V_i)] + \mathbb{E}\Big[\mathbb{E}[\alpha_i \mid V_1, \ldots, V_{i-1}]\Big| \; V_i, \mathrm{mp}_{\mathcal{G}}(V_i)\Big]
\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}[\alpha_i \mid \text{mp}_{\mathcal{G}}(V_i)] - \mathbb{E}\Big[\mathbb{E}[\alpha_i \mid V_1, \ldots, V_{i-1}]\Big|\ \text{mp}_{\mathcal{G}}(V_i)\Big] \\
&= \alpha_i - \mathbb{E}[\alpha_i \mid V_1, \ldots, V_{i-1}] - \mathbb{E}[\alpha_i \mid V_i, \text{mp}_{\mathcal{G}}(V_i)] + \mathbb{E}[\alpha_i \mid \text{mp}_{\mathcal{G}}(V_i)] \\
&= \Big\{\alpha_i - \mathbb{E}[\alpha_i \mid V_1, \ldots, V_{i-1}]\Big\} - \Big\{\mathbb{E}\Big[\alpha_i - \mathbb{E}[\alpha_i \mid V_i, \ldots, V_{i-1}] \ \Big|\ V_i, \text{mp}_{\mathcal{G}}(V_i)\Big]\Big\}.
\end{aligned}
$$

Consequently, the orthogonal complement of the tangent space is the following,

$$
\Lambda^{*\perp} = \Big\{ \sum_{V_i \in V} \alpha_i(V_1, \ldots, V_i) - \mathbb{E}[\alpha_i \mid V_i, \text{mp}_{\mathcal{G}}(V_i)] \Big\},
$$

where α_i is any function of $V_1, \ldots, V_i$ such that $\mathbb{E}[\alpha_i \mid V_1, \ldots, V_{i-1}] = 0$, i.e., $\alpha_i \in \Lambda_i$. ■

Lemma 4 (Primal IPW formulation)

Proof Our goal is to demonstrate that the primal IPW formulation is equivalent to the identifying functional of the target parameter $\psi(t)$ shown in Eq. 14 and restated below.

$$
\psi(t) = \sum_{V \setminus T} \prod_{V_i \in V \setminus D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\Big|_{T=t} \times \sum_{T} \prod_{D_i \in D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y.
$$

The primal IPW formulation for the target $\psi(t)$ is,

$$
\mathbb{E}[\beta_{\text{primal}}(t)] \equiv \mathbb{E}\Big[\frac{\mathbb{I}(T=t)}{q_{D_T}(T \mid \text{mb}_{\mathcal{G}}(T))} \times Y \ \Big],
$$

where $q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T)) = \prod_{V_i \in D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))$, and

$$
\begin{aligned}
q_{D_T}(T \mid \text{mb}_{\mathcal{G}}(T)) &= q_{D_T}(T \mid D_T \cup \text{pa}_{\mathcal{G}}(D_T) \setminus T) = \frac{q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))}{q_{D_T}(D_T \setminus T \mid \text{pa}_{\mathcal{G}}(D_T))} \\
&= \frac{q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))}{\sum_T q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))} = \frac{\prod_{V_i \in D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))}{\sum_T \prod_{V_i \in D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))} \\
&= \frac{\prod_{V_i \in \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))}{\sum_T \prod_{V_i \in \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))}.
\end{aligned}
$$

The last equality holds because the conditional densities of $V_i \in \mathbb{C}$, does not depend on T, and they cancel out from the numerator and denominator. Therefore, product in the ratio is over the variables in $D_T \cap \{\succeq T\}$ which we have denoted by $\mathbb{L}$. Therefore,

$$
\begin{aligned}
\mathbb{E}[\beta_{\text{primal}}(t)] &= \mathbb{E}\Big[\mathbb{I}(T=t) \times \frac{\sum_T \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))}{\prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))} \times Y \ \Big] \\
&= \sum_V \prod_{V_i \in V} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \mathbb{I}(T=t) \times \frac{\sum_T \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))}{\prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))} \times Y \\
&= \sum_V \mathbb{I}(T=t) \times \prod_{V_i \in V \setminus \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \\
&\qquad \times \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times \frac{\sum_T \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))}{\prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))} \times Y
\end{aligned}
$$

$$= \sum_{V} \mathbb{I}(T = t) \times \prod_{V_i \in V \setminus \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \sum_{T} \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y.$$

In the second equality, we evaluated the outer expectation with respect to the joint $p(V)$. In the third equality, we partitioned the joint into factors for the set $\mathbb{L}$ and factors for $V \setminus \mathbb{L}$. In the fourth equality, we canceled out the factors involved in the denominator of the primal IPW with the corresponding terms in the joint.

We can then move the conditional factors of pre-treatment variables in the district of T past the summation over T as these factors are not functions of T. Finally, we evaluate the indicator function, concluding the proof. That is,

$$\psi_{\text{primal}} = \sum_{V} \mathbb{I}(T = t) \times \prod_{V_i \in V \setminus D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \sum_{T} \prod_{D_i \in D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y$$

$$= \sum_{V \setminus T} \prod_{V_i \in V \setminus D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\Big|_{T=t} \times \sum_{T} \prod_{D_i \in D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y = \psi(t)$$

■

Lemma 5 (Dual IPW formulation)

Proof The proof strategy is similar to the one used for the primal IPW. The dual IPW formulation for the target $\psi(t)$ is,

$$\begin{aligned}
\mathbb{E}[\beta_{\text{dual}}(t)] &= \mathbb{E}\Big[\frac{\prod_{M_i \in \mathbb{M}^*} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\mid_{T=t}}{\prod_{M_i \in \mathbb{M}^*} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times Y\Big] \\
&= \sum_{V} \prod_{V_i \in V} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \frac{\prod_{M_i \in \mathbb{M}^*} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\mid_{T=t}}{\prod_{M_i \in \mathbb{M}^*} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times Y \\
&= \sum_{V} \prod_{V_i \in V \setminus \mathbb{M}^*} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \\
&\qquad \times \prod_{M_i \in \mathbb{M}^*} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \times \frac{\prod_{M_i \in \mathbb{M}^*} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\mid_{T=t}}{\prod_{M_i \in \mathbb{M}^*} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times Y \\
&= \sum_{V} \prod_{V_i \in V \setminus \mathbb{M}^*} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \prod_{M_i \in \mathbb{M}^*} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\mid_{T=t} \times Y \\
&= \sum_{V \setminus T} \prod_{V_i \in V \setminus \{\mathbb{M}^* \cup D_T\}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \prod_{M_i \in \mathbb{M}^*} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\mid_{T=t} \\
&\qquad \times \sum_{T} \prod_{D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y.
\end{aligned}$$

In the above derivation, we first evaluated the outer expectation with respect to the joint $p(V)$. We then partitioned the joint into factors corresponding to $\mathbb{M}^*$ and $V \setminus \mathbb{M}^*$. The factors involved in the denominator of the dual IPW then canceled out with the corresponding terms in the joint. The last equality holds because by the definition of the inverse Markov pillow, $\mathbb{M}^*$ contains all variables not in the district of T such that T is a member of its Markov pillow. In the above expression, factors corresponding to the inverse Markov pillow of T are

evaluated at $T = t$. Consequently, the only factors above that are still functions of T are the ones corresponding to the district of T. This allows us to push the summation over T.

Finally, since the summation over T will prevent factors within the district of T from being evaluated at $T = t$, we can simply apply the evaluation to the entire functional and merge the sets not involved in the district of T above. That is,

$$\psi_{\text{dual}} = \sum_{V \setminus T} \prod_{V_i \in V \setminus D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\Big|_{T=t} \times \sum_{T} \prod_{D_i \in D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y = \psi(t).$$

■

Theorem 6 (Primal and Dual IPW estimators)

Proof The required positivity assumptions for $\widehat{\psi(t)}_{\text{primal}}$ and $\widehat{\psi(t)}_{\text{dual}}$ are as follows.

$$\widehat{\psi(t)}_{\text{primal}}: \quad \forall L_i \in \mathbb{L},\ p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) > \epsilon_{l_i} \ \text{ almost surely for some non-negative } \epsilon_{l_i},$$
$$\widehat{\psi(t)}_{\text{dual}}: \quad \forall M_i \in \mathbb{M}^*,\ p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) > \epsilon_{m_i} \ \text{ almost surely for some non-negative } \epsilon_{m_i}.$$

For the usual regularity conditions see Theorem 1A in Robins et al. (1992). Under these conditions, the proof of asymptotic normality is fairly standard once we show the corresponding estimating equations are unbiased. In Lemmas (4) and (5), we proved the unbiasedness of these two estimators, i.e., we showed $\mathbb{E}[\beta(t)_{\text{primal}}] = \mathbb{E}[\beta(t)_{\text{dual}}] = \psi(t)$.

For the sake of completeness, we walk through finding the asymptotic variance. Let $U(\psi, \eta)$ denote either $\beta(t)_{\text{primal}} - \psi(t)$ or $\beta(t)_{\text{dual}} - \psi(t)$. ($\eta$ denotes the set of nuisance parameters.) Given our proof in Lemmas (4) and (5), we know that $\mathbb{E}[U(\psi_0, \eta_0)] = 0$. In order to estimate ψ_0, we use the estimating equation: $\frac{1}{n} \times \sum_{i=1}^{n} U_i\big(\psi(\hat{\eta}), \hat{\eta}\big) = 0$. The Taylor series expansion of $U_i\big(\psi(\hat{\eta}), \hat{\eta}\big)$ around ψ_0 is

$$\begin{aligned}
0 =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i\Big(\psi(\hat{\eta}), \hat{\eta}\Big) \\
=& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i\Big(\psi_0, \hat{\eta}\Big) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \psi} U_i\Big(\psi_0, \hat{\eta}\Big) \times (\psi(\hat{\eta}) - \psi_0) + o_p(1) \\
=& \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i\Big(\psi_0, \hat{\eta}\Big)}_{(a)} + \underbrace{\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \psi} U_i\Big(\psi_0, \hat{\eta}\Big)}_{(b)} \times \sqrt{n}(\psi(\hat{\eta}) - \psi_0) + o_p(1)
\end{aligned}$$

$$\begin{aligned}
(a) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i\Big(\psi_0, \hat{\eta}\Big) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i\Big(\psi_0, \eta_0\Big) + \frac{1}{n} \sum_{i=1}^{n} \frac{\partial U_i(\psi_0, \eta_0)}{\partial \eta} \times \sqrt{n}(\widehat{\eta} - \eta_0) + o_p(1)
\end{aligned}$$

$$\text{as } n \to \infty \ (a) \to \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i\Big(\psi_0, \eta_0\Big) + \mathbb{E}\left[\frac{\partial U(\psi_0, \eta_0)}{\partial \eta}\right] \times \sqrt{n}(\widehat{\eta} - \eta_0)$$

$$
\begin{aligned}
(b) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \psi}\, U_i\Big(\psi_0, \hat{\eta}\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \psi}\Big\{U_i\Big(\psi_0, \eta_0\Big) + \frac{\partial}{\partial \eta} U_i(\psi_0, \eta_0) \times (\widehat{\eta} - \eta_0)\Big\} \\
\text{as } n \to \infty \ (b) &\to \mathbb{E}\Big[\frac{\partial}{\partial \psi} U\Big(\psi_0, \eta_0\Big)\Big]
\end{aligned}
$$

By substituting (a) and (b), we get:

$$
0 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\, U_i\Big(\psi_0, \eta_0\Big) + \mathbb{E}\left[\frac{\partial U(\psi_0, \eta_0)}{\partial \eta}\right] \times \sqrt{n}(\widehat{\eta} - \eta_0) + \mathbb{E}\Big[\frac{\partial}{\partial \psi} U\Big(\psi_0, \eta_0\Big)\Big] \times \sqrt{n}\Big(\psi(\hat{\eta}) - \psi_0\Big) + o_p(1).
$$

Consequently,

$$
\sqrt{n}\Big(\psi(\hat{\eta}) - \psi_0\Big) = -\mathbb{E}\Big[\frac{\partial}{\partial \psi} U\Big(\psi_0, \eta_0\Big)\Big]^{-1} \times \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\, U_i\Big(\psi_0, \eta_0\Big) + \mathbb{E}\left[\frac{\partial U(\psi_0, \eta_0)}{\partial \eta}\right] \times \sqrt{n}(\widehat{\eta} - \eta_0)\right) + o_p(1).
$$

The regularity conditions guarantee that,

$$
\sqrt{n}(\widehat{\eta} - \eta_0) = -\mathbb{E}\Big[\frac{\partial S}{\partial \eta}\Big]^{-1} \times \frac{1}{\sqrt{n}} S(\eta_0) + o_p(1).
$$

$$
\begin{aligned}
\sqrt{n}\Big(\psi(\hat{\eta}) - \psi_0\Big) &= -\mathbb{E}\Big[\frac{\partial}{\partial \psi} U\Big(\psi_0, \eta_0\Big)\Big]^{-1} \times \frac{1}{\sqrt{n}} \times \left(\sum_{i=1}^{n}\, U_i\Big(\psi_0, \eta_0\Big) - \mathbb{E}\left[\frac{\partial U(\psi_0, \eta_0)}{\partial \eta}\right] \times \mathbb{E}\Big[\frac{\partial S}{\partial \eta}\Big]^{-1} \times S(\eta_0)\right) + o_p(1). \\
&= -\tau^{-1} \times \frac{1}{\sqrt{n}} \times \sum_{i}\, \text{Resid}(U_i) + o_p(1).
\end{aligned}
$$

As a consequence of central limit theorem and Slutzky's theorem, we conclude that $\sqrt{n}\big(\psi(\hat{\eta}) - \psi_0\big)$ is asymptotically normal with mean zero and variance

$$
\text{Var}(\psi) = \text{Var}(\tau^{-1}\text{Resid}) = \tau^{-1} \times \text{Var}(\text{Resid}) \times \tau^{-1},
$$

that can be consistently estimated using the sample variance of $\text{Var}(\psi)$.

■

Theorem 7 (Variational independence of primal IPW and dual IPW)

Proof Consider the topological factorization of the observed distribution $p(V)$ for the ADMG as shown in Eq. 7.

$$
p(V) = \prod_{V_i \in V} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)).
$$

Note by definition, the inverse Markov pillow of T does not contain elements in the district of T, i.e., $\mathbb{M}^* \cap D_T = \emptyset$. Thus, we can partition V into three disjoint sets as follows:

$$\mathbb{L} = D_T \cap \{\succeq T\}, \qquad \mathbb{M}^*, \qquad \mathbb{R} = V \setminus (\mathbb{L} \cup \mathbb{M}^*)$$

The set $\mathbb{L}$ is the same as what we introduced in Section 4. $\mathbb{M}^*$ is a subset of $\mathbb{M}$, and the remaining variables are in set $\mathbb{R} = \mathbb{C} \cup (\mathbb{M} \setminus \mathbb{M}^*)$. The topological factorization of the observed joint can then be restated as,

$$p(V) = \prod_{R_i \in \mathbb{R}} p(R_i \mid \text{mp}_{\mathcal{G}}(R_i)) \prod_{M_i \in \mathbb{M}^*} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)).$$

It is then clear from the above factorization that the components of the primal IPW estimator which sit in $\mathbb{L}$, and the components of the dual IPW estimator which sit in $\mathbb{M}^*$, form congenial and variationally independent pieces of the joint distribution $p(V)$. ■

Theorem 8 (The efficient influence function of $\psi(t)$ in $\mathcal{M}_{\text{np}}$)

Proof As a reminder, we partition the set of nodes V into three disjoint sets: $V = \{\mathbb{C}, \mathbb{L}, \mathbb{M}\}$, where

$$\mathbb{C} = \{C_i \in V \mid C_i \prec T\}, \quad \mathbb{L} = \{L_i \in V \mid L_i \in D_T, L_i \succeq T\}, \text{ and } \mathbb{M} = V \setminus \mathbb{C} \cup \mathbb{L}.$$

The target parameter is identified via the following function of the observed data,

$$\psi_\kappa(t) = \sum_{V \setminus T} Y \times \prod_{M_i \in \mathbb{M}} p_\kappa(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \sum_{T} \prod_{L_i \in \mathbb{L}} p_\kappa(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times p_\kappa(\mathbb{C}), \tag{41}$$

and according to Eq. 35, $\frac{d}{d\kappa}\psi_\kappa(t)\big|_{\kappa=0} = \mathbb{E}[U_{\psi_t} \times S_{\eta_0}(V)]$. Therefore,

$$\begin{aligned}
\frac{d}{d\kappa}\psi_\kappa(t) &= \frac{d}{d\kappa}\Big\{ \sum_{V \setminus T} Y \times \prod_{M_i \in \mathbb{M}} p_\kappa(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \sum_{T} \prod_{L_i \in \mathbb{L} \setminus T} p_\kappa(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times p_\kappa(T, \mathbb{C}) \Big\} \\
&= \sum_{V \setminus T} Y \times \frac{d}{d\kappa}\Big\{ \prod_{M_i \in \mathbb{M}} p_\kappa(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \Big\} \times \sum_{T} \prod_{L_i \in \mathbb{L} \setminus T} p_\kappa(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times p_\kappa(T, \mathbb{C}) \quad \text{(1st Term)} \\
&+ \sum_{V \setminus T} Y \times \prod_{M_i \in \mathbb{M}} p_\kappa(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \sum_{T} \frac{d}{d\kappa}\Big\{ \prod_{L_i \in \mathbb{L} \setminus T} p_\kappa(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \Big\} \times p_\kappa(T, \mathbb{C}) \quad \text{(2nd Term)} \\
&+ \sum_{V \setminus T} Y \times \prod_{M_i \in \mathbb{M}} p_\kappa(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \sum_{T} \prod_{L_i \in \mathbb{L} \setminus T} p_\kappa(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \frac{d}{d\kappa}\Big\{ p_\kappa(T, \mathbb{C}) \Big\}. \quad \text{(3rd Term)}
\end{aligned}$$

First Term: The contribution of the first term to the final IF is made of individual contributions of the elements in $\mathbb{M}$. Since the derivation is similar, we only derive it for an element $M_j \in \mathbb{M}$.

$$\begin{aligned}
&\sum_{V \setminus T} Y \times \prod_{M_i \in \{\prec M_j\} \cap \mathbb{M}} p_\kappa(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \frac{d}{d\kappa}\Big\{ p_\kappa(M_j \mid \text{mp}_{\mathcal{G}}(M_j))|_{T=t} \Big\} \\
&\qquad \times \prod_{M_i \in \{\succ M_j\} \cap \mathbb{M}} p_\kappa(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \sum_{T} \prod_{L_i \in \mathbb{L}} p_\kappa(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times p_\kappa(\mathbb{C})
\end{aligned}$$

$$
\begin{aligned}
&\overset{(1)}{=} \sum_{V\setminus\{T,\{\preceq M_j\}\}} \prod_{M_i\in\{\prec M_j\}\cap\mathbb{M}} p_\kappa(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \frac{d}{d\kappa}\Big\{p_\kappa(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j))|_{T=t}\Big\} \\
&\qquad \times \sum_{T\cup\{\succ M_j\}} Y \times \prod_{V_i\in\mathbb{L}\cup\{\{\succ M_j\}\cap\mathbb{M}\}} p_\kappa(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))\Big|_{T=t \text{ if } V_i\in\mathbb{M}} \times p_\kappa(\mathbb{C}) \\
&\overset{(2)}{=} \sum_{\preceq M_j} \frac{\mathbb{I}(T=t)}{\prod_{L_i\prec M_j} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times S(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j)) \times \prod_{V_i\in\{\preceq M_j\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \\
&\qquad \times \sum_{T\cup\{\succ M_j\}} Y \times \prod_{V_i\in\mathbb{L}\cup\{\{\succ M_j\}\cap\mathbb{M}\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))\Big|_{T=t \text{ if } V_i\in\mathbb{M}} \\
&\overset{(3)}{=} \mathbb{E}\Bigg[\frac{\mathbb{I}(T=t)}{\prod_{L_i\prec M_j} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \underbrace{\sum_{T\cup\{\succ M_j\}} Y \times \prod_{V_i\in\mathbb{L}\cup\{\{\succ M_j\}\cap\mathbb{M}\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))\Big|_{T=t \text{ if } V_i\in\mathbb{M}}}_{:=f(\preceq M_j)} \\
&\qquad\qquad \times S(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j))\Bigg] \\
&\overset{(4)}{=} \mathbb{E}\Bigg[\frac{\mathbb{I}(T=t)}{\prod_{L_i\prec M_j} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \Big(f(\preceq M_j) - \sum_{M_j} f(\preceq M_j) \times p(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j))\Big) \times S(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j))\Bigg] \\
&\overset{(5)}{=} \mathbb{E}\Bigg[\frac{\mathbb{I}(T=t)}{\prod_{L_i\prec M_j} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \Big(f(\preceq M_j) - \sum_{M_j} f(\preceq M_j) \times p(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j))\Big) \times S(V)\Bigg]
\end{aligned}
$$

The first equality follows from the fact that terms corresponding to $M_i \in \{\prec M_j\}$ are not functions of elements in $\{\succ M_j\}$ and of Y. The second equality follows by term grouping, the definition of conditional scores, and term cancellation. The third equality is by definition of joint expectation. The fourth and fifth equalities are implied by the fact that conditional scores have expected value of 0 (given their conditioning set). Therefore, the contribution of $M_j \in \mathbb{M}$ is the following:

$$
\begin{aligned}
&\frac{\mathbb{I}(T=t)}{\prod_{L_i\prec M_j} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \Big(\sum_{T\cup\{\succ M_j\}} Y \times \prod_{V_i\in\mathbb{L}\cup\{\{\succ M_j\}\cap\mathbb{M}\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))\Big|_{T=t \text{ if } V_i\in\mathbb{M}} \\
&\qquad\qquad - \sum_{T\cup\{\succeq M_j\}} Y \times \prod_{V_i\in\mathbb{L}\cup\{\{\succeq M_j\}\cap\mathbb{M}\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))\Big|_{T=t \text{ if } V_i\in\mathbb{M}} \Big).
\end{aligned}
$$

Second Term: The contribution of the second term to the final IF is made of individual contributions of the elements in $\mathbb{L}\setminus T$. Since the derivation is similar, we only derive it for an element $L_j \in \mathbb{L}\setminus T$.

$$
\begin{aligned}
&\sum_{V\setminus T} Y \times \prod_{M_i\in\mathbb{M}} p_\kappa(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \sum_{T} \Bigg\{ \prod_{L_i\in\{\prec L_j\}\cap\mathbb{L}\setminus T} p_\kappa(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \\
&\qquad \times \frac{d}{d\kappa}\Big\{p_\kappa(L_j \mid \mathrm{mp}_{\mathcal{G}}(L_j))\Big\} \times \prod_{L_i\in\{\succ L_j\}\cap\mathbb{L}\setminus T} p_\kappa(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))\Bigg\} \times p_\kappa(T,\mathbb{C}) \\
&\overset{(1)}{=} \sum_{V} Y \times \prod_{V_i\in\{\succ L_j\}} p_\kappa(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))|_{T=t \text{ if } V_i\in\mathbb{M}} \times \frac{d}{d\kappa}\Big\{p_\kappa(L_j \mid \mathrm{mp}_{\mathcal{G}}(L_j))\Big\} \\
&\qquad\qquad \times \prod_{V_i\in\{\prec L_j\}} p_\kappa(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))|_{T=t \text{ if } V_i\in\mathbb{M}} \\
&\overset{(2)}{=} \sum_{\preceq L_j} \underbrace{\sum_{\succ L_j} Y \times \prod_{V_i\in\{\succ L_j\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))|_{T=t \text{ if } V_i\in\mathbb{M}}}_{f(\preceq L_j)} \times S(L_j \mid \mathrm{mp}_{\mathcal{G}}(L_j))
\end{aligned}
$$

$$\times \prod_{V_i \in \{\preceq L_j\}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\big|_{T=t \text{ if } V_i \in \mathbb{M}}$$

$$\overset{(3)}{=} \sum_{\preceq L_j} f(\preceq L_j) \times \frac{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\big|_{T=t}}{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times S(L_j \mid \text{mp}_{\mathcal{G}}(L_j)) \times \prod_{V_i \in \{\preceq L_j\}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))$$

$$\overset{(4)}{=} \mathbb{E}\left[\frac{\prod_{M_i \in \mathbb{M}\{\cap \prec L_j\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\big|_{T=t}}{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times f(\preceq L_j) \times S(L_j \mid \text{mp}_{\mathcal{G}}(L_j))\right]$$

$$\overset{(5)}{=} \mathbb{E}\left[\frac{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\big|_{T=t}}{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times \Big(f(\preceq L_j) - \sum_{L_j} f(\preceq L_j) \times p(L_j \mid \text{mp}_{\mathcal{G}}(L_j))\Big) \times S(L_j \mid \text{mp}_{\mathcal{G}}(L_j))\right]$$

$$\overset{(6)}{=} \mathbb{E}\left[\frac{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\big|_{T=t}}{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times \Big(f(\preceq L_j) - \sum_{L_j} f(\preceq L_j) \times p(L_j \mid \text{mp}_{\mathcal{G}}(L_j))\Big) \times S(V)\right]$$

The first equality follows from the fact that terms corresponding to $M_i \in \mathbb{M}$ are not functions of T, the fact that $\mathbb{C}, \mathbb{M}, \mathbb{L}$ partition V, and term grouping. The second equality is by definition of conditional scores. The third equality is by term cancellation. The fourth is by definition of joint expectations, the fifth and sixth equalities are implied by the fact that conditional scores have expected value of 0 (given their conditioning set). Therefore, the contribution of $L_j \in \mathbb{L} \setminus T$ is the following:

$$\frac{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\big|_{T=t}}{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times \Big(\sum_{\succ L_j} Y \times \prod_{V_i \in \{\succ L_j\}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\big|_{T=t \text{ if } V_i \in \mathbb{M}} - \sum_{\succeq L_j} Y \times \prod_{V_i \in \{\succeq L_j\}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\big|_{T=t \text{ if } V_i \in \mathbb{M}} \Big).$$

Third Term: The contribution of the last term to the final IF is as follows.

$$\sum_{V \setminus T} Y \times \prod_{M_i \in \mathbb{M}} p_\kappa(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \sum_{T} \prod_{L_i \in \mathbb{L} \setminus T} p_\kappa(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \frac{d}{d\kappa}\Big\{p_\kappa(T, \mathbb{C})\Big\}$$

$$\overset{(1)}{=} \sum_{T, \mathbb{C}} \Bigg\{ \underbrace{\sum_{V \setminus T, \mathbb{C}} Y \times \prod_{M_i \in \mathbb{M}} p_\kappa(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \prod_{L_i \in \mathbb{L} \setminus T} p_\kappa(L_i \mid \text{mp}_{\mathcal{G}}(L_i))}_{f(T, \mathbb{C})} \Bigg\} \times \frac{d}{d\kappa}\Big\{p_\kappa(T, \mathbb{C})\Big\}.$$

$$\overset{(2)}{=} \sum_{T, \mathbb{C}} f(T, \mathbb{C}) \times S(T, \mathbb{C}) \times p(T, \mathbb{C}) = \mathbb{E}\Big[f(T, \mathbb{C}) \times S(T, \mathbb{C})\Big]$$

$$\overset{(3)}{=} \mathbb{E}\Big[\Big(f(T, \mathbb{C}) - \sum_{T, \mathbb{C}} f(T, \mathbb{C}) \times p(T, \mathbb{C})\Big) \times S(T, \mathbb{C})\Big]$$

$$\overset{(4)}{=} \mathbb{E}\Big[\Big(f(T, \mathbb{C}) - \psi(t)\Big) \times S(V)\Big].$$

The first equality is term grouping, the second is by definition of marginal scores, the third and fourth equalities are implied by the fact that scores have expected value 0. Therefore, the contribution of the last term is the following:

$$\sum_{V \setminus \{T, \mathbb{C}\}} Y \times \prod_{M_i \in \mathbb{M}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\big|_{T=t} \times \prod_{L_i \in \mathbb{L} \setminus T} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) - \psi(t).$$

Putting all these together yields the final influence function. ■

Theorem 9 (Double robustness of augmented primal IPW)

Proof We need to show that under correct specification of conditional densities in either $\{p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)), \forall M_i \in \mathbb{M}\}$ or $\{p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)), \forall L_i \in \mathbb{L}\}$, the influence function in Theorem 8 remains to be mean zero. We break this down into two scenarios.

Scenario 1. Assume models in $\mathbb{L}$ are correctly specified, and let $p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))$ denote the misspecified model for $p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)), \forall M_i \in \mathbb{M}$. We note that for any $L_j \in \mathbb{L} \setminus T$, the following line in the IF evaluates to zero in expectation.

$$
\begin{aligned}
&\mathbb{E}\Big[\frac{\prod_{M_i \prec L_j} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \Big(\sum_{\succ L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succ L_j\}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \\
&\qquad - \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq L_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \Big) \Big] \\
&\overset{(1)}{=} \sum_{\preceq L_j} \frac{\prod_{M_i \prec L_j} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times \prod_{V_i \prec L_j} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times p(L_j \times \text{mp}_{\mathcal{G}}(L_j)) \\
&\qquad \times \Big(\sum_{\succ L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succ L_j\}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \\
&\qquad\qquad - \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \Big) \\
&\overset{(2)}{=} \sum_{\prec L_j} \frac{\prod_{M_i \prec L_j} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times \prod_{V_i \prec L_j} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \sum_{L_j} p(L_j \times \text{mp}_{\mathcal{G}}(L_j)) \\
&\qquad \times \Big(\sum_{\succ L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succ L_j\}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \\
&\qquad\qquad - \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \Big) \\
&\overset{(3)}{=} \sum_{\prec L_i} \frac{\prod_{M_i \prec L_j} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times \prod_{V_i \prec L_j} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \\
&\qquad \times \Big(\sum_{\succ L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succ L_j\}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \\
&\qquad\qquad - \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \Big) \\
&\overset{(4)}{=} 0.
\end{aligned}
$$

The first equality is by definition of joint expectation. The second equality is by the fact that terms associated with $\prec L_j$ are not functions of L_j. The third equality is by term grouping.

Moreover, for any $M_j, M_{j-1} \in \mathbb{M}$, the following equality holds,

$$
\begin{aligned}
&\mathbb{E}\Big[\frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \sum_{T \cup \{\succeq M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big] \\
&= \mathbb{E}\Big[\frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_{j-1}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \sum_{T \cup \{\succ M_{j-1}\}} Y \times \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ M_{j-1}\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big],
\end{aligned}
$$

since the left hand side is equal to

$$\sum_{\prec M_i} p(\prec M_i) \times \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \Big[\sum_{T \cup \{\succeq M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big]$$

$$\overset{(1)}{=} \sum_{\preceq M_{j-1}} p(\preceq M_{i-1}) \times \Bigg\{ \sum_{M_{j-1} \prec L_k \prec M_j} \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times p(L_k \mid \text{mp}_{\mathcal{G}}(L_k)) \times \Big[\sum_{T \cup \{\succeq M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big] \Bigg\}$$

$$\overset{(2)}{=} \sum_{\preceq M_{j-1}} p(\preceq M_{j-1}) \times \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_{j-1}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \sum_{M_{j-1} \prec L_k \prec M_j} \Big\{ \sum_{T \cup \{\succeq M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big\}$$

$$\overset{(3)}{=} \sum_{\preceq M_{j-1}} p(\preceq M_{j-1}) \times \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_{j-1}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \Big\{ \sum_{T \cup \{\succ M_{j-1}\}} Y \times \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big\}$$

$$\overset{(4)}{=} \mathbb{E}\Bigg[\frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_{j-1}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \Big\{ \sum_{T \cup \{\succ M_{j-1}\}} Y \times \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \{\mathbb{M} \cap \succ M_{j-1}\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big\} \Bigg],$$

which is exactly the same as the right hand side. This leaves the IF with only two terms $\psi(t)$ and β_{primal} and according to Lemma 4, $\mathbb{E}[\beta_{\text{primal}}] = \psi(t)$, provided the models in $\mathbb{L}$ are correctly specified, which was assumed. Therefore, $\mathbb{E}[U_{\psi_t}] = 0$.

Scenario 2. Assume models in $\mathbb{M}$ are correctly specified, and let $p^*(L_i \mid \text{mp}_{\mathcal{G}}(L_i))$ denote the misspecified model for $p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)), \forall L_i \in \mathbb{L}$. We note that for any $M_j \in \mathbb{M}$, the following line in the IF evaluates to zero.

$$\mathbb{E}\Big[\frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p^*(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \Big(\sum_{T \cup \{\succ M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p^*(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ M_j\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} - \sum_{T \cup \{\succeq M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p^*(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \Big) \Big]$$

$$\overset{(1)}{=} \sum_{\preceq M_j} \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p^*(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \prod_{V_i \prec M_j} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times p(M_j \mid \text{mp}_{\mathcal{G}}(M_j)) \times \Big(\sum_{T \cup \{\succ M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p^*(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ M_j\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} - \sum_{T \cup \{\succeq M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p^*(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \Big)$$

$$\overset{(2)}{=} \sum_{\prec M_j} \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p^*(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \prod_{V_i \prec M_j} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \sum_{M_j} p(M_j \mid \text{mp}_{\mathcal{G}}(M_j))$$

$$\times \Big(\sum_{T \cup \{\succ M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ M_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}$$

$$- \sum_{T \cup \{\succeq M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \Big)$$

$$\stackrel{(3)}{=} \sum_{\prec M_j} \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \prod_{V_i \prec M_j} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))$$

$$\times \Big(\sum_{T \cup \{\succeq M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}$$

$$- \sum_{T \cup \{\succeq M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \Big)$$

$$\stackrel{(4)}{=} 0.$$

Moreover, for any $L_j, L_{j-1} \in \mathbb{L}$, the following equality holds,

$$\mathbb{E}\Big[\frac{\prod_{M_i \prec L_j} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))}$$

$$\times \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big]$$

$$\mathbb{E}\Big[\frac{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))}$$

$$\times \sum_{\succ L_{j-1}} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succ L_{j-1}\}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_{j-1}\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big],$$

since the left hand side is equal to

$$\sum_{\prec L_j} p(\prec L_j) \times \frac{\prod_{M_i \prec L_j} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))}$$

$$\times \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \mid_{T=t}$$

$$\stackrel{(1)}{=} \sum_{\preceq L_{j-1}} p(\preceq L_{j-1}) \times \Big\{ \sum_{L_{j-1} \prec M_k \prec L_j} \frac{\prod_{M_i \prec L_j} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times p(M_k \mid \mathrm{mp}_{\mathcal{G}}(M_k))$$

$$\times \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big\}$$

$$\stackrel{(2)}{=} \sum_{\preceq L_{j-1}} p(\preceq L_{j-1}) \times \frac{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times \Big\{ \sum_{L_{j-1} \prec M_k \prec L_j} p(M_k \mid \mathrm{mp}_{\mathcal{G}}(M_k))|_{T=t}$$

$$\times \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big\} \Big\}$$

$$\stackrel{(3)}{=} \sum_{\preceq L_{j-1}} p(\preceq L_{j-1}) \times \frac{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times$$

$$\times \Big\{ \sum_{L_{j-1} \prec M_k \prec L_j} \Big[\sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big]$$

$$\times\, p(M_k \mid \text{mp}_{\mathcal{G}}(M_k))|_{T=t}\Big\}$$

$$\overset{(4)}{=} \sum_{\preceq L_{j-1}} p(\preceq L_{j-1}) \times \frac{\prod_{M_i \prec L_{j-1}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_{j-1}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))}$$

$$\times \sum_{\succ L_{j-1}} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succ L_{j-1}\}} p^*(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_{j-1}\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\ |_{T=t}\Big\}$$

$$\overset{(5)}{=} \mathbb{E}\Bigg[\frac{\prod_{M_i \prec L_{j-1}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_{j-1}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))}$$

$$\times \sum_{\succ L_{j-1}} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succ L_{j-1}\}} p^*(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_{j-1}\}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\ |_{T=t}\Bigg],$$

which is exactly the same as the right hand side. This leaves the IF with only two terms $\psi(t)$ and β_{dual} and according to Lemma 5, $\mathbb{E}[\beta_{\text{dual}}] = \psi(t)$. Therefore, $\mathbb{E}[U_{\psi_t}] = 0$. ■

Lemma 10 (Reformulation of the IF for augmented primal IPW)

Proof We prove this lemma by showing what happens to $V_i \in V$, if V_i is in $\mathbb{M}$, or $\mathbb{L}$, or $\mathbb{C}$.

∘ For any $M_j \in \mathbb{M}$, we have,

$$\mathbb{E}\Big[\beta_{\text{primal}} \;\Big|\; \{\preceq M_j\}\Big] = \mathbb{E}\Bigg[\frac{\mathbb{I}(T=t)}{\prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \sum_T \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times Y \;\Bigg|\; \{\preceq M_i\}\Bigg]$$

$$= \sum_{V_i \succ M_j} \prod_{V_i \succ M_j} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \frac{\mathbb{I}(T=t)}{\prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \sum_T \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times Y$$

$$= \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \sum_{V_i \succ M_j} \sum_T \prod_{V_i \in \mathbb{L} \cup \{\succ M_j\}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))|_{T=t \text{ if } V_i \in \mathbb{M}} \times Y\Big\}$$

$$= \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \sum_{T \cup \{\succ M_j\}} Y \times \prod_{V_i \in \mathbb{L} \cup \{\succ M_j\}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))|_{T=t \text{ if } V_i \in \mathbb{M}}\Big\}.$$

Similarly,

$$\mathbb{E}\big[\beta_{\text{primal}} \mid \{\prec M_j\}\big] = \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \sum_{T \cup \{\succeq M_j\}} Y \times \prod_{V_i \in \mathbb{L} \cup \{\succeq M_j\}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))|_{T=t \text{ if } V_i \in \mathbb{M}}\Big\}.$$

Therefore, $\mathbb{E}\big[\beta_{\text{primal}} \mid \{\preceq M_j\}\big] - \mathbb{E}\big[\beta_{\text{primal}} \mid \prec \{M_j\}\big]$ is equivalent to M_j's corresponding line in the IF.

∘ Now, for any $L_j \in \mathbb{L}$, we have,

$$\mathbb{E}\Big[\beta_{\text{dual}} \;\Big|\; \{\preceq L_j\}\Big] = \mathbb{E}\Bigg[\frac{\prod_{M_i \in \mathbb{M}}\ p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \in \mathbb{M}}\ p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times Y \;\Bigg|\; \{\preceq L_j\}\Bigg]$$

$$= \sum_{V_i \succ L_j} \prod_{V_i \succ L_j} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \frac{\prod_{M_i \in \mathbb{M}}\ p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \in \mathbb{M}}\ p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times Y$$

$$= \frac{\prod_{M_i \prec L_j}\ p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j}\ p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times \sum_{V_i \succ L_j} Y \times \prod_{V_i \succ L_j} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))|_{T=t \text{ if } V_i \in \mathbb{M}}.$$

Similarly,

$$\mathbb{E}\Big[\,\beta_{\text{dual}}\;\Big|\;\{\prec L_j\}\Big] = \frac{\prod_{M_i \prec L_j}\; p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j}\; p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times \sum_{V_i \succeq L_j} Y \times \prod_{V_i \succeq L_j} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))|_{T=t \text{ if } V_i \in \mathbb{M}}.$$

Therefore, $\mathbb{E}\big[\,\beta_{\text{dual}} \mid \{\preceq L_j\}\big] - \mathbb{E}\big[\,\beta_{\text{dual}} \mid \{\prec L_j\}\big]$ is equivalent to L_j's corresponding line in the IF.

$\circ$ For variables in $\mathbb{C}$, we have,

$$\begin{aligned}
\mathbb{E}[\beta_{\text{primal}} \mid \mathbb{C}] &= \mathbb{E}\Big[\frac{\mathbb{I}(T=t)}{\prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \sum_{T} \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times Y \;\Big|\; \mathbb{C}\Big] \\
&= \sum_{V \setminus \mathbb{C}} p(V \setminus \mathbb{C}) \times \frac{\mathbb{I}(T=t)}{\prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times \sum_{T} \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times Y \\
&= \sum_{V \setminus \mathbb{C}} \mathbb{I}(T=t) \times \prod_{M_i \in \mathbb{M}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \times \sum_{T} \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times Y \\
&= \sum_{V \setminus \{T, \mathbb{C}\}} \prod_{M_i \in \mathbb{M}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\Big|_{T=t} \times \sum_{T} \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times Y,
\end{aligned}$$

and based on Lemma 4, $\mathbb{E}[\beta_{\text{primal}}] = \psi(t)$. Therefore, $\mathbb{E}[\beta_{\text{primal}} \mid \mathbb{C}] - \mathbb{E}[\beta_{\text{primal}}]$ corresponds to the last line in the IF. We can also run a similar argument for β_{dual}. According to Lemma 5, $\mathbb{E}[\beta_{\text{dual}}] = \psi(t)$, and

$$\begin{aligned}
\mathbb{E}[\beta_{\text{dual}} \mid \mathbb{C}] &= \mathbb{E}\Big[\frac{\prod_{M_i \in \mathbb{M}}\; p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \in \mathbb{M}}\; p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times Y \;\Big|\; \mathbb{C}\Big] \\
&= \sum_{V \setminus \mathbb{C}} p(V \setminus \mathbb{C}) \times \frac{\prod_{M_i \in \mathbb{M}}\; p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \in \mathbb{M}}\; p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times Y \\
&= \sum_{V \setminus \mathbb{C}} \prod_{M_i \in \mathbb{M}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times Y \\
&= \sum_{V \setminus \{T, \mathbb{C}\}} \prod_{M_i \in \mathbb{M}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\Big|_{T=t} \times \sum_{T} \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times Y \\
&= \mathbb{E}[\beta_{\text{primal}} \mid \mathbb{C}].
\end{aligned}$$

■

Lemma 11 (Identifying functional when T is fixable)

Proof A set Z satisfies the backdoor criterion (and so yields identification of the targe via the adjustment functional) with respect to a treatment T and outcome Y if (i) Z does not contain any descendants of T; and (ii) T and Y are m-separated by Z in a graph $\mathcal{G}_{\bar{T}}$ where all outgoing edges $T \to \circ$ from the treatment are removed (Pearl, 2009). We now show that $\text{mp}_{\mathcal{G}}(T)$ satisfies both criteria. By the pre-condition that $\text{dis}_{\mathcal{G}}(T) \cap \text{de}_{\mathcal{G}}(T) = \{T\}$, there exists a valid topological ordering on vertices V such that T appears last among the members of its district. Under such an ordering $\text{mp}_{\mathcal{G}}(T) = \text{mb}_{\mathcal{G}}(T)$. That is, under fixability

$\text{mp}_{\mathcal{G}}(T) = \text{dis}_{\mathcal{G}}(T) \cup \text{pa}_{\mathcal{G}}(\text{dis}_{\mathcal{G}}(T)) \setminus T$. To see (i) is satisfied we note that $\text{mp}_{\mathcal{G}}(T)$ does not contain any descendants of T. To see (ii) is satisfied, we first note that Y is a non-descendant of T in $\mathcal{G}_{\bar{T}}$, and the Markov blanket of T in this graph remains the same and excludes the outcome Y when T is fixable. Further, from Richardson et al. (2017) we know that the Markov blanket of a variable m-separates the variable from all its non-descendants (excluding the Markov blanket itself.) From this it follows that $T \perp\!\!\!\perp Y \mid \text{mp}_{\mathcal{G}}(T)$ in $\mathcal{G}_{\bar{T}}$. ■

Theorem 12 (Efficient augmented primal IPW in mb-shielded ADMGs)

Proof Consider the reformulated IF in Lemma 10. In order to get the efficient IF, we project the reformulated IF onto the tangent space Λ^* given by Lemma 3. We first note that we can rewrite the term $\sum_{\mathbb{C}} \mathbb{E}[\beta_{\text{primal/dual}} \mid \mathbb{C}] - \psi(t)$ in the reformulated IF as $\sum_{C_i \in \mathbb{C}} \mathbb{E}[\beta_{\text{primal/dual}} \mid \{\preceq C_i\}] - \mathbb{E}[\beta_{\text{primal/dual}} \mid \{\prec C_i\}]$, where $\beta_{\text{primal/dual}}$ means that we can use either β_{primal} or β_{dual} for the $\mathbb{C}$ term. We have,

$$\begin{aligned}
\pi[U^{\text{reform}}_{\psi_t} \mid \Lambda^*] = &\sum_{M_i \in \mathbb{M}} \pi\Big[\mathbb{E}[\beta_{\text{primal}} \mid \{\preceq M_i\}] - \mathbb{E}[\beta_{\text{primal}} \mid \{\prec M_i\}]\Big| \ \Lambda^*\Big] \\
&+ \sum_{L_i \in \mathbb{L}} \pi\Big[\mathbb{E}[\beta_{\text{dual}} \mid \{\preceq L_i\}] - \mathbb{E}[\beta_{\text{dual}} \mid \{\prec L_i\}] \ \Big| \ \Lambda^*\Big] \\
&+ \sum_{C_i \in \mathbb{C}} \pi\Big[\mathbb{E}[\beta_{\text{primal/dual}} \mid \preceq C_i] - \mathbb{E}[\beta_{\text{primal/dual}} \mid \prec C_i] \ \Big| \ \Lambda^*\Big].
\end{aligned}$$

Let β be either β_{primal} or β_{dual} or $\beta_{\text{primal/dual}}$. Note that $\Big\{\mathbb{E}\big[\beta \mid \{\preceq V_i\}\big] - \mathbb{E}\big[\beta_{\text{primal}} \mid \{\prec V_i\}\big]\Big\}$ lives in Λ_{V_i}, and $\Lambda_{V_i} \perp\!\!\!\perp \Lambda^* \setminus \Lambda^*_{V_i}$. Therefore, their projection onto $\Lambda^* \setminus \Lambda^*_{V_i}$ is zero. We have,

$$\begin{aligned}
&\pi\Big[\mathbb{E}\big[\beta \mid \{\preceq V_i\}\big] - \mathbb{E}\big[\beta \mid \{\prec V_i\}\big]\Big| \ \Lambda^*_{V_i}\Big] \\
&\quad = \mathbb{E}\Big[\mathbb{E}[\beta \mid \{\preceq V_i\}] - \mathbb{E}[\beta \mid \{\prec V_i\}] \ \Big| \ V_i, \text{mp}_{\mathcal{G}}(V_i)\Big] - \mathbb{E}\Big[\mathbb{E}[\beta \mid \{\preceq V_i\}] - \mathbb{E}[\beta \mid \{\prec V_i\}] \ \Big| \ \text{mp}_{\mathcal{G}}(V_i)\Big] \\
&\quad = \mathbb{E}\big[\beta \mid V_i, \text{mp}_{\mathcal{G}}(V_i)\big] - \mathbb{E}\Big[\mathbb{E}\big[\beta \mid \prec V_i\big] \ \Big| \ V_i, \text{mp}_{\mathcal{G}}(V_i)\Big] - \mathbb{E}\big[\beta \mid \ \text{mp}_{\mathcal{G}}(V_i)\big] + \mathbb{E}\big[\beta \mid \ \text{mp}_{\mathcal{G}}(V_i)\big] \\
&\quad = \mathbb{E}\big[\beta \mid V_i, \text{mp}_{\mathcal{G}}(V_i)\big] - \mathbb{E}\Big[\mathbb{E}\big[\beta \mid \prec V_i\big] \ \Big| \ V_i, \text{mp}_{\mathcal{G}}(V_i)\Big] \\
&\quad = \mathbb{E}\big[\beta \mid V_i, \text{mp}_{\mathcal{G}}(V_i)\big] - \mathbb{E}\big[\beta \mid \ \text{mp}_{\mathcal{G}}(V_i)\big].
\end{aligned}$$

Therefore, the efficient IF is as follows.

$$\begin{aligned}
\pi[U^{\text{reform}}_{\psi_t} \mid \Lambda^*] = &\sum_{M_i \in \mathbb{M}} \mathbb{E}\big[\beta_{\text{primal}} \mid M_i, \text{mp}_{\mathcal{G}}(M_i)\big] - \mathbb{E}\big[\beta_{\text{primal}} \mid \text{mp}_{\mathcal{G}}(M_i)\big] \\
&+ \sum_{L_i \in \mathbb{L}} \mathbb{E}\big[\beta_{\text{dual}} \mid L_i, \text{mp}_{\mathcal{G}}(L_i)\big] - \mathbb{E}\big[\beta_{\text{dual}} \mid \text{mp}_{\mathcal{G}}(L_i)\big] \\
&+ \sum_{C_i \in \mathbb{C}} \mathbb{E}\big[\beta_{\text{primal/dual}} \mid C_i, \text{mp}_{\mathcal{G}}(C_i)\big] - \mathbb{E}\big[\beta_{\text{primal/dual}} \mid \text{mp}_{\mathcal{G}}(C_i)\big].
\end{aligned}$$

■

Lemma 13 (Efficient augmented IPW in mb-shielded ADMGs)

Proof An mb-shielded ADMG is Markov equivalent to a DAG $\mathcal{G}^d$, which can be constructed as follows. Under the topological order τ fixed on the original ADMG $\mathcal{G}$, $V_i \to V_j$ exists in $\mathcal{G}^d$ if V_i and V_j are adjacent in $\mathcal{G}$ and $V_i \prec_\tau V_j$. $\mathcal{G}^d$ is a DAG because we only allow for directed edges and there is no directed cycle as we follow a valid topological order in $\mathcal{G}$. Further, $\text{mp}_\mathcal{G}(V_i) = \text{pa}_{\mathcal{G}^d}(V_i), \forall V_i \in V$. Therefore, the identifying functional for the target parameter is the same in both $\mathcal{G}$ and $\mathcal{G}^d$, that is $\mathbb{E}[\mathbb{E}[Y \mid T = t, \text{mp}_\mathcal{G}(T)] = \mathbb{E}[\mathbb{E}[Y \mid T = t, \text{pa}_{\mathcal{G}^d}(T)]]$.

We know for the instrumental variables in

$$Z = \{Z_i \in V \mid Z_i \perp\!\!\!\perp Y \mid \text{mp}_\mathcal{G}(Z_i) \text{ in } \mathcal{G}_{V \setminus T} \text{ and } Z_i \not\perp\!\!\!\perp T \mid \text{mp}_\mathcal{G}(Z_i)\},$$

there always exists a set $F \in V$ that d-separates $Z_i \in Z$ from Y given F, T (van der Zander et al., 2015; Rotnitzky and Smucler, 2020). Showing that the equation

$$\mathbb{E}\Big[\frac{\mathbb{I}(T=t)}{p(T \mid \text{mp}_\mathcal{G}(T))} \times Y \;\Big|\; Z_i, \text{mp}_\mathcal{G}(Z_i)\Big] = \mathbb{E}\Big[\frac{\mathbb{I}(T=t)}{p(T \mid \text{mp}_\mathcal{G}(T))} \times Y \;\Big|\; \text{mp}_\mathcal{G}(Z_i)\Big].$$

holds then simply follows from the argument outlined in Proposition 3 of Rotnitzky and Smucler (2020).
Finally, given Λ^* in Lemma 3, the efficient IF is as follows. Let $V^* = V \setminus (T \cup Z \cup D)$,

$$\begin{aligned} U^{\text{eff}}_{\psi_t} &= \pi\Big[\frac{\mathbb{I}(T=t)}{p(T \mid \text{mp}_\mathcal{G}(T))} \times Y \;\Big|\; \Lambda^* \setminus \Lambda^*_T\Big] \\ &= \sum_{V_i \in V^*} \mathbb{E}\Big[\frac{\mathbb{I}(T=t)}{p(T \mid \text{mp}_\mathcal{G}(T))} \times Y \;\Big|\; V_i, \text{mp}_\mathcal{G}(V_i)\Big] - \mathbb{E}\Big[\frac{\mathbb{I}(T=t)}{p(T \mid \text{mp}_\mathcal{G}(T))} \times Y \;\Big|\; \text{mp}_\mathcal{G}(V_i)\Big]. \end{aligned}$$

■

Theorem 14 (Soundness and completeness of Algorithm 2)

Proof Soundness of the algorithm implies that when our algorithm succeeds, the subsequent identifying functional for $\psi(t)$ is correct. Completeness implies, that when the algorithm fails, the target parameter $\psi(t)$ is not identifiable within the model.

Soundness

We first prove soundness of the algorithm. That is, when Algorithm 2 does not fail, $\psi(t)$ is indeed equal to $\psi(t)_{\text{nested}}$. The algorithm does not fail when all districts $D \in \mathcal{D}^*$ are intrinsic in $\mathcal{G}$. Note that $\mathcal{D}^*$ is a subset of the districts in $\mathcal{G}_{Y^*}$. However, by construction of $\mathcal{D}^*$, the remaining districts in $\mathcal{G}_{Y^*}$ are those that do not have any overlap with D_T. We now show that such districts are always intrinsic in $\mathcal{G}$.

Consider a district $D \in \mathcal{D}(\mathcal{G}_{Y^*})$ such that $D \cap D_T = \emptyset$. The district D forms a subset of a larger district in $\mathcal{G}$, say $D' \in \mathcal{D}(\mathcal{G})$. Due to results in (Tian and Pearl, 2002a), we know

that D' is always intrinsic. If $D = D'$ then the result immediately follows. Otherwise, In the CADMG $\phi_{V \setminus D'}(\mathcal{G})$, there exists at least one vertex D_i in D' not in Y^*, that has no children. This is because all directed paths from D_i to vertices in Y^* must go through T and since T is not in D', all incoming edges to T have been deleted. The only other way D_i may not be childless is if there existed a cycle in $\mathcal{G}$, which is a contradiction. Thus, such a vertex D_i is always fixable and furthermore, fixing it corresponds to the marginalization operation $\sum_{D_i} q_{D'}(D' \mid \text{pa}_{\mathcal{G}}(D'))$ (Richardson et al., 2017). Once D_i is fixed, another vertex D_j that is in D' but not in Y^* becomes childless. Applying this argument inductively, we see that all $D_i \in D'$ such that $D_i \notin Y^*$ are fixable through marginalization under a reverse topological order. Hence for districts D in $\mathcal{G}_{Y^*}$ that do not overlap with D_T, the set $D = D' \setminus \{D_i \in D' \mid D_i \notin Y^*\}$ is always intrinsic. Thus, Algorithm 2 succeeds when all districts in $\mathcal{G}_{Y^*}$ are intrinsic.

Soundness

We now show that under this condition, $\psi(t)_{\text{nested}} \equiv \mathbb{E}_{p^\dagger}\left[\frac{\mathbb{I}(T=t)}{p(T|\text{mp}_{\mathcal{G}}(T))} \times Y\right] = \psi(t)$. By definition, we have

$$\psi(t)_{\text{nested}} = \sum_V p(V) \times \prod_{D^* \in \mathcal{D}^*} \frac{q_{D^*}(D^* \mid \text{pa}_{\mathcal{G}}(D^*))}{\prod_{D_i^* \in D^*} p(D_i^* \mid \text{mp}_{\mathcal{G}}(D_i^*))} \times \frac{\mathbb{I}(T=t)}{p(T \mid \text{mp}_{\mathcal{G}}(T))} \times Y.$$

The districts of $\mathcal{G}$ can be partitioned into three sets. $\mathcal{D}_T$ is the district in $\mathcal{G}$ that contains T (with all elements in $\mathcal{D}^*$, if any, subsets of D_T). $\mathcal{D}'$ is the set of districts in $\mathcal{G}$, excluding D_T, that overlap with Y^*. $\mathcal{D}^z$ is the set of districts in $\mathcal{G}$, excluding D_T, that do not overlap with Y^*. The observed distribution $p(V)$ then district factorizes as,

$$p(V) = \prod_{D^z \in \mathcal{D}^z} q_{D^z}(D^z \mid \text{pa}_{\mathcal{G}}(D^z)) \times \prod_{D' \in \mathcal{D}'} q_{D'}(D' \mid \text{pa}_{\mathcal{G}}(D')) \times q_{D_T}(D_T | \text{pa}_{\mathcal{G}}(D_T)).$$

By results in Tian and Pearl (2002a), $q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))$ is identified as $\prod_{D_i \in D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))$ (for any topological ordering). Since every element in $\mathcal{D}^*$ is a subset of D_T, and since vertices in $D_T \setminus \bigcup_{D^* \in \mathcal{D}^*}$ precede vertices $D_T \cap \bigcup_{D^* \in \mathcal{D}^*} = D_T \cap Y^*$ in the ordering, we have

$$\begin{aligned}\psi(t)_{\text{nested}} = \sum_V \prod_{D^z \in \mathcal{D}^z} q_{D^z}(D^z \mid \text{pa}_{\mathcal{G}}(D^z)) \times \prod_{D' \in \mathcal{D}'} q_{D'}(D' \mid \text{pa}_{\mathcal{G}}(D')) \times \prod_{D^* \in \mathcal{D}^*} q_{D^*}(D^* \mid \text{pa}_{\mathcal{G}}(D^*)) \\ \times \sum_{D_T \cap Y^*} q_{D_T}(D_T | \text{pa}_{\mathcal{G}}(D_T)) \times \frac{\mathbb{I}(T=t)}{p(T \mid \text{mp}_{\mathcal{G}}(T))} \times Y.\end{aligned}$$

Since T is the last element in the ordering in $D_T \setminus Y^*$, we further have:

$$\begin{aligned}\psi(t)_{\text{nested}} = \sum_{Y^*} \sum_{V \setminus Y^*} \prod_{D^z \in \mathcal{D}^z} q_{D^z}(D^z \mid \text{pa}_{\mathcal{G}}(D^z)) \times \prod_{D' \in \mathcal{D}'} q_{D'}(D' \mid \text{pa}_{\mathcal{G}}(D')) \times \prod_{D^* \in \mathcal{D}^*} q_{D^*}(D^* \mid \text{pa}_{\mathcal{G}}(D^*)) \\ \times \sum_{(D_T \cap Y^*) \cup \{T\}} q_{D_T}(D_T | \text{pa}_{\mathcal{G}}(D_T)) \times \mathbb{I}(T=t) \times Y.\end{aligned}$$

Consider applying marginalization of elements in $V \setminus Y^*$ to $\psi(t)_{\text{nested}}$ above in the reverse topological ordering on $V \setminus Y^*$. Districts in $\mathcal{G}$ partition V and so, by definition of $\mathcal{D}^*, \mathcal{D}'$ and

D_T, elements in $\mathcal{D}^z \cup \{D' \setminus Y^* : D' \in \mathcal{D}'\} \cup \{D_T \setminus (Y^* \cup \{T\})\}$ partition $V \setminus Y^*$. This partition, and the fact that marginalizations are processed in reverse topological order, means that at every stage, the variable to be summed occurs in precisely one place in the expression. This implies that the result of the overall summation of $V \setminus Y^*$ yields:

$$\psi(t)_{\text{nested}} = \sum_{Y^*} \prod_{D' \in \mathcal{D}'} \sum_{D' \setminus Y^*} q_{D'}(D' \mid \text{pa}_{\mathcal{G}}(D')) \times \prod_{D^* \in \mathcal{D}^*} q_{D^*}(D^* \mid \text{pa}_{\mathcal{G}}(D^*)) \times \mathbb{I}(T = t) \times Y$$

By definition, $q_{D^*}(D^* \mid \text{pa}_{\mathcal{G}}(D^*)) \equiv \phi_{V \setminus D^*}(p(V); \mathcal{G}(V))$. Since every D' in $\mathcal{D}'$ is a top level district in $\mathcal{G}$, there exists a valid fixing sequence on $V \setminus D'$. Further, in the CADMG $\phi_{V \setminus D'}(\mathcal{G}(V))$, any element in $D' \setminus Y^*$ cannot be an ancestor of an element in $D' \cap Y^*$ (if a directed path not through T existed from an element V_i in D' to an element in $D' \cap Y^*$, then V_i must itself be in $D' \cap Y^*$, while a directed path from V_i to $D' \cap Y^*$ through T disappears in $\phi_{V \setminus D'}(\mathcal{G}(V))$ since T is outside D'. Consequently fixing elements $D' \setminus Y^*$ in reverse topological order in $\phi_{V \setminus D'}(\mathcal{G}(V))$ and $\phi_{V \setminus D'}(p(V), \mathcal{G}(V))$ is equivalent to marginalizing those variables. As a result, for every $D' \in \mathcal{D}'$, $\sum_{D' \setminus Y^*} q_{D'}(D' \mid \text{pa}_{\mathcal{G}}(D')) = \phi_{V \setminus (D' \cap Y^*)}(p(V); \mathcal{G}(V))$. Our conclusion follows:

$$\psi(t)_{\text{nested}} = \sum_{Y^*} \prod_{D \in \mathcal{D}(\mathcal{G}_{Y^*})} \phi_{V \setminus D}(p(V); \mathcal{G}) \times Y \Big|_{T=t} = \psi(t).$$

Completeness

Follows trivially as we have shown the failure condition of Algorithm 2 to be equivalent to the failure condition of the identification algorithm in Richardson et al. (2017) which is known to be sound and complete. ■

Lemma 15 (Commutativity of p-fixing)

Proof Consider a valid p-fixing sequence $(S_1, \ldots, S_p)$ for the set S. That the kernel $\Phi_{(S_1,\ldots,S_p)}(p(V); \mathcal{G})$ evaluated at any $s_1, \ldots, s_p$ is equal to $p(V \setminus \{S_1, \ldots, S_p\} \mid \text{do}(s_1, \ldots, s_p))$ follows by an inductive application of Theorem 3 in Tian and Pearl (2002a). That is, for any two valid p-fixing sequences σ_S^1 and σ_S^2 defined on S we have that $\Phi_{\sigma_S^1}(p(V); \mathcal{G}) = \Phi_{\sigma_S^2}(p(V); \mathcal{G}) = p(V \setminus S \mid \text{do}(S = s))$. ■

Corollary 16 (Identification via a sequence of p-fixing)

Proof Given any valid p-fixing sequence $(Z_1, \ldots, Z_p, T)$ the kernel $\Phi_{(Z_1,\ldots,Z_p,T)}(p(V); \mathcal{G})$ evaluated at any $z_1, \ldots, z_p, t$ is equal to $p(V \setminus \{Z_1, \ldots, Z_p, T\} \mid \text{do}(z_1, \ldots, z_p, t))$. Then $p(Y \mid \text{do}(z_1, \ldots, z_p, t)) = \sum_{V \setminus \{Z_1,\ldots,Z_p,T,Y\}} p(V \setminus \{Z_1, \ldots, Z_p, T\} \mid \text{do}(z_1, \ldots, z_p, t))$. Since all vertices $Z_1, \ldots, Z_p$ have no directed paths to Y except through T in the original graph, the corresponding exclusion restrictions in the causal model imply $p(Y \mid \text{do}(z_1, \ldots, z_p, t)) = p(Y \mid \text{do}(t))$. ■

References

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

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

Peter J. Bickel, Chris A.J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Marco Carone, Alexander R. Luedtke, and Mark J. van der Laan. Toward computerized efficient estimation in infinite-dimensional models. *Journal of the American Statistical Association*, 114(527):1174–1190, 2019.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney K. Newey, and James M. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21, 2018.

George B. Dantzig, Lester R. Ford Jr., and Delbert R. Fulkerson. A primal-dual algorithm. *Linear Equalities and Related Systems, Annals of Mathematics Study*, 38:171–181, 1956.

Mathias Drton. Discrete chain graph models. *Bernoulli*, 15(3):736–753, 2009.

Robin J. Evans. Margins of discrete Bayesian networks. *The Annals of Statistics*, 46(6A): 2623–2656, 2018.

Robin J. Evans and Thomas S. Richardson. Smooth, identifiable supermodels of discrete DAG models with latent variables. *Bernoulli*, 25(2):848–876, 2019.

Constantine E. Frangakis, Tianchen Qian, Zhenke Wu, and Iván Díaz. Deductive derivation and Turing-computerization of semiparametric efficient estimation. *Biometrics*, 71(4): 867–874, 2015.

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

Sander Greenland, Judea Pearl, and James M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, pages 37–48, 1999.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

Jaroslav Hájek. A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(4):323–330, 1970.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Leonard Henckel, Emilija Perković, and Marloes H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 2021.

Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 13–16, 2006.

Paul Hünermund and Elias Bareinboim. Causal inference and data-fusion in econometrics. *arXiv preprint arXiv:1912.09104*, 2019.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating causal effects using weighting-based estimators. In *Proceedings of the 34th Conference on Artificial Intelligence*. AAAI Press, 2020.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating identifiable causal effects through double machine learning. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2021.

Edward H. Kennedy, Zongming Ma, Matthew D. McHugh, and Dylan S. Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245, 2017.

Steffen L. Lauritzen. *Graphical Models*. Oxford, U.K.: Clarendon, 1996.

David G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, 1997.

Daniel Malinsky, Ilya Shpitser, and Thomas S. Richardson. A potential outcomes calculus for identifying conditional path-specific effects. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 3080–3088, 2019.

Maxima. Maxima: a computer algebra system, 2020. URL `http://maxima.sourceforge.net/`.

Whitney K. Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5 (2):99–135, 1990.

Ryo Okui, Dylan S. Small, Zhiqiang Tan, and James M. Robins. Doubly robust instrumental variable regression. *Statistica Sinica*, pages 173–205, 2012.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. A complete generalized adjustment criterion. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 682–691. AUAI Press, 2015.

Thomas S. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

Thomas S. Richardson and James M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128 (30), 2013.

Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*, 2017.

J. Robins, L. Li, E. Tchetgen, and A. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. *Collections*, 2:335–421, 2008.

James M. Robins. 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, 1986.

James M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, The Environment, and Clinical Trials*, pages 95–133. Springer, 2000.

James M Robins, Steven D Mark, and Whitney K Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, pages 479–495, 1992.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

A. Rotnitzky, E. Smucler, and J.M. Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108:231–238, 2020.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models. *Journal of Machine Learning Research*, 2020.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006.

Ilya Shpitser, Tyler J. VanderWeele, and James M. Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 527–536. AUAI Press, 2010.

Ilya Shpitser, Robin J. Evans, and Thomas S. Richardson. Acyclic linear SEMs obey the nested Markov property. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 2018.

Ilya Shpitser, Thomas S. Richardson, and James M. Robins. Multivariate counterfactual systems and causal graphical models. `https://arxiv.org/abs/2008.06017`, 2020.

Peter L. Spirtes, Clark N. Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas S. Richardson. *Causation, prediction, and search*. MIT press, 2000.

Masashi Sugiyama, Motoaki Kawanabe, and Pui Ling Chui. Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, 23(1):44–59, 2010.

Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2):146–160, 1972.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence*, pages 567–573. American Association for Artificial Intelligence, 2002a.

Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 519–527, 2002b.

Santtu Tikka and Juha Karvanen. Simplifying probabilistic expressions in causal inference. *Journal of Machine Learning Research*, 18(1):1203–1232, 2017.

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

Mark J. van der Laan and James M. Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

Mark J. van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

Aad W. van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Benito van der Zander, Johannes Textor, and Maciej Liśkiewicz. Efficiently finding conditional instruments for causal inference. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 2015.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, 1990.

Linbo Wang and Eric J. Tchetgen Tchetgen. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society Series B, Statistical Methodology*, 80(3):531, 2018.

Thomas C. Williams, Cathrine C. Bach, Niels B. Matthiesen, Tine B. Henriksen, and Luigi Gagliardi. Directed acyclic graphs: a tool for causal studies in paediatrics. *Pediatric Research*, 84(4):487–493, 2018.