
A Characterization of the Orthocomplement of the Tangent Space of Semiparametric Markov Models

Trung Phung¹

Ilya Shpitser¹

¹Computer Science Department, Johns Hopkins University, Baltimore, Maryland, USA

Abstract

Graphical models are ubiquitous in social and empirical science as they are intuitive and easy to use. These models belong to the broader class of Markov models, defined using solely conditional independence (CI) restrictions.

In order to estimate finite-dimensional target parameters in such models efficiently, semiparametric theory provides a principled framework for constructing regular and asymptotically linear estimators via influence functions (IFs). These estimators are asymptotically normal and root- n consistent. Characterizing the class of all influence functions for a target parameter is crucial for statistically efficient inference in these models.

For models that are Markov relative to directed acyclic graphs (DAGs), the orthogonal complement of the tangent space is known, implying that for any target the class of all influence functions can be derived once an influence function is obtained. On the other hand, for Markov models not equivalent to a DAG model – such as ordinary Markov models associated with undirected graphs, chain graphs, or acyclic directed mixed graphs – the orthogonal complement has not been characterized, impeding semi-parametric inference in these models.

We derive closed form expressions for the orthogonal complement of the tangent space for general Markov models and illustrate our results by characterizing the class of influence functions for the conditional mean parameter in several graphical models.

1 INTRODUCTION

A common task in statistical and causal inference is estimating a finite dimensional target parameter from data. In parametric models where elements are indexed by finite dimensional nuisance parameters, efficient estimation is achieved using the theory of maximum likelihood estimation [van der Vaart, 2000]. In nonparametric or semiparametric models where the parameterization is infinite dimensional, regular and asymptotically linear (RAL) estimators are desirable, since they have nice properties such as asymptotic normality and root- n rate of convergence. RAL estimators are typically obtained by deriving influence function of the target parameter [Newey, 1990, Bickel et al., 1993, Van Der Vaart, 2002, Tsiatis, 2006, Kennedy, 2024].

The class of influence functions of a parameter of interest in a given semiparametric model are elements of the linear variety constructed using one influence function and the orthogonal complement of the tangent space of the model (Tsiatis [2006], Theorem 4.3). Characterizing the class of all possible RAL estimators in a given semiparametric model thus entails characterizing the orthogonal complement of the tangent space of the model.

Statistical independence is an important tool for encoding *irrelevance*, and is thus crucial for obtaining interpretable models for empirical phenomena where relationships among variables are local. Thus, in this paper, we consider Markov models, or conditional independence models, which are statistical models defined using solely marginal and conditional independences, with a focus on the subclass of Markov models associated with graphs, referred to as *graphical Markov models* [Pearl, 1988, Lauritzen, 1996, Maathuis et al., 2018].

Graphical models have been associated with directed acyclic graphs (DAGs), undirected graphs (UGs), as well as mixed graphs such as chain graphs (CGs) [Frydenberg, 1990] and acyclic directed mixed graphs (ADMGs) [Pearl, 1988, Lauritzen, 1996, Richardson, 2003, Maathuis

et al., 2018]. Models associated with DAGs have seen wide use due to their interpretability, and connections to causal models [Pearl, 2000, Spirtes et al., 2001, Richardson and Robins, 2013]. However, other types of graphical models have also seen uses for associational analysis of gene regulatory networks [Segal et al., 2003], protein signaling networks [Sachs et al., 2005], network data [Tchetgen Tchetgen et al., 2021, Ogburn et al., 2020], and spatial data [Lopes et al., 2008], as well as modeling systems at equilibrium [Lauritzen and Richardson, 2002]. In addition, ADMG models have been connected to causal models with hidden variables, as well as theory of identification in such models [Richardson, 2003, Richardson et al., 2023, Shpitser et al., 2022, Shpitser and Pearl, 2006].

Explicit likelihoods in terms of sets of variationally independent components have been derived for DAG models, leading to a natural characterization of their model tangent spaces in terms of orthogonal subspaces [Tsiatis, 2006, Rotnitzky and Smucler, 2020, Bhattacharya et al., 2022]. While such likelihoods are also known for UG and CG models [Shpitser, 2023], they feature overlapping components, meaning their tangent space cannot easily be expressed in terms of orthogonal subspaces. Furthermore, no general likelihoods for ADMG models are known. Since these Markov models associated with UGs, CGs, and ADMGs are not, in general, observationally equivalent to models associated with DAGs, this greatly complicates the derivation of the tangent space and hence its orthogonal complement, necessitating the development of new techniques.

In this paper, we derive a characterization of the orthogonal complement of the tangent space for semiparametric Markov models defined by marginal and conditional independences, including graphical models associated with undirected graphs, chain graphs, and acyclic directed mixed graphs. In addition, we illustrate how our characterization may be used to derive all influence functions for target parameters arising in statistical and causal inference applications, with the conditional mean parameter as a worked example.

2 PRELIMINARIES

Prior to discussing our contribution, we first review the necessary preliminaries: conditional independence (Markov) models and the important subclass of graphical Markov models, as well as the theory of statistical inference in semiparametric models.

2.1 MARKOV MODELS

Suppose $p(\mathbf{V})$ is a distribution over the set of all random variables of interest $\mathbf{V}$. If $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are disjoint subsets of $\mathbf{V}$, a *conditional independence* (CI) statement “ $\mathbf{X}$ and $\mathbf{Y}$

are independent given $\mathbf{Z}$ in p ”, denoted $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$, means $p(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{y} \mid \mathbf{z})$ for all values of $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ such that $p(\mathbf{z}) > 0$. Given a list $\mathcal{L}$ of CI statements, a statistical model associated with these constraints, written as $\{p(\mathbf{V}) : \text{all CIs in } \mathcal{L} \text{ hold in } p(\mathbf{V})\}$ is called a *Markov model*, or a *conditional independence model* [Maathuis et al., 2018]. Important subclasses of Markov models are *graphical models* associated with directed acyclic graphs (DAGs), undirected graphs (UGs), chain graphs (CGs), and acyclic directed mixed graph (ADMGs) [Maathuis et al., 2018, Lauritzen, 1996, Richardson, 2003].

Undirected graphs (UGs) are graphs with only undirected edges ($-$). Directed acyclic graphs (DAGs) are graphs with only directed edges ($\rightarrow$) lacking directed cycles. Bidirected graphs (BGs) are graphs with only bidirected edges ($\leftrightarrow$). Chain graphs (CGs) are mixed graphs containing both directed and undirected edges with no partially directed cycles. Acyclic directed mixed graphs (ADMGs) are mixed graphs with directed and bidirected edges lacking directed cycles (for the purposes of such cycles the presence of bidirected edges is ignored). Simple examples of these graphs classes are shown in Fig. 1.

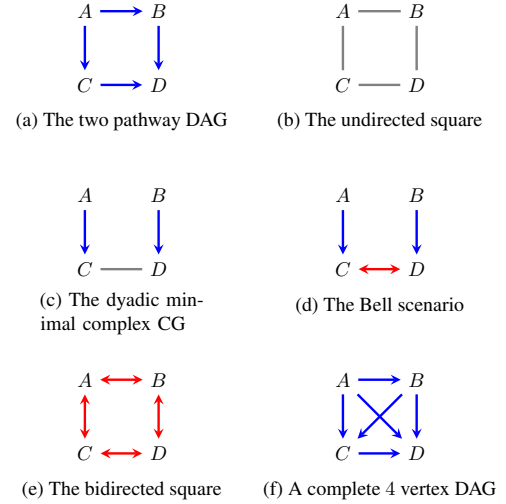


Figure 1: Six different graphs used to define six different graphical Markov models. Models associated with graphs in (b), (c), (d), and (e) are not equivalent to any DAG model, so their tangent spaces cannot be characterized using the technique in Section 3.

Graphical models use graphical separation criteria to encode CI constraints. The (finite) set of all CIs that hold in a graphical model is given by a global Markov property, while local Markov properties provide a small list of CIs that logically imply all others. Below, we define local Markov properties for models associated with UGs, DAGs, CGs, bidirected graphs (BGs), and ADMGs, and refer reader to Maathuis et al. [2018], Richardson [2003] for additional details.

To describe the local Markov properties, we need the following basic graphical concepts. If $\mathcal{G}$ is an UG, denote $\text{ne}_\mathcal{G}(V) \equiv \{Z \in \mathbf{V} : Z - V \text{ in } \mathcal{G}\}$ as the set of neighbors of V in $\mathcal{G}$. If $\mathcal{G}$ is a DAG, denote $\text{pa}_\mathcal{G}(V) \equiv \{Z \in \mathbf{V} : Z \rightarrow V \text{ in } \mathcal{G}\}$ as the set of parents of V in $\mathcal{G}$, $\text{ch}_\mathcal{G}(V) \equiv \{Z \in \mathbf{V} : V \rightarrow Z \text{ in } \mathcal{G}\}$ as the set of children of V in $\mathcal{G}$, $\text{deg}_\mathcal{G}(V) \equiv \{Z \in \mathbf{V} : V \rightarrow \cdots \rightarrow Z \text{ in } \mathcal{G}\}$ as the set of descendants of V in $\mathcal{G}$, $\text{ang}_\mathcal{G}(V) \equiv \{Z \in \mathbf{V} : Z \rightarrow \cdots \rightarrow V \text{ in } \mathcal{G}\}$ as the set of ancestors of V in $\mathcal{G}$, and $\text{nd}_\mathcal{G}(V) \equiv \mathbf{V} \setminus \text{deg}_\mathcal{G}(V)$ as the set of non-descendants of V in $\mathcal{G}$. Note that by convention, $V \in \text{deg}_\mathcal{G}(V) \cap \text{ang}_\mathcal{G}(V)$. For sets of variables, the sets of parents, ancestors and descendants are defined disjunctively. That is $\text{pa}_\mathcal{G}(\mathbf{A}) \equiv \bigcup_{A \in \mathbf{A}} \text{pa}_\mathcal{G}(A)$, $\text{ang}_\mathcal{G}(\mathbf{A}) \equiv \bigcup_{A \in \mathbf{A}} \text{ang}_\mathcal{G}(A)$ and $\text{deg}_\mathcal{G}(\mathbf{A}) \equiv \bigcup_{A \in \mathbf{A}} \text{deg}_\mathcal{G}(A)$.

For CGs, define $\text{pa}_\mathcal{G}(V)$ as in DAGs and $\text{ne}_\mathcal{G}(V)$ as in UGs. The set of descendants of V , $\text{deg}_\mathcal{G}(V)$ is defined as the set of all variables Z with a *partially directed path* from V to Z – a sequence of distinct vertices such that there is an edge ($-$ or $\rightarrow$) between any pair of consecutive vertices, with all directed edges pointing in the same direction. As before, $V \in \text{deg}_\mathcal{G}(V)$ and $\text{nd}_\mathcal{G}(V) = \mathbf{V} \setminus \text{deg}_\mathcal{G}(V)$. We further define the *boundary of V* , $\text{bd}_\mathcal{G}(V)$ as $\text{ne}_\mathcal{G}(V) \cup \text{pa}_\mathcal{G}(V)$. Finally, we define a chain component in a CG $\mathcal{G}$ to be a connected component in the edge subgraph of $\mathcal{G}$ retaining only undirected edges.

If $\mathcal{G}$ contains bidirected edges, denote $\text{sib}_\mathcal{G}(V) \equiv \{Z \in \mathbf{V} : Z \leftrightarrow V \text{ in } \mathcal{G}\}$ as the set of siblings of V in $\mathcal{G}$. Furthermore, a subset $\mathbf{C} \subseteq \mathbf{V}$ is called a bidirected connected subset if it is a connected component in the edge subgraph of $\mathcal{G}$ retaining only bidirected edges. A district is a maximal bidirected connected subset, and $\text{dis}_\mathcal{G}(V)$ denotes the district in $\mathcal{G}$ containing V .

For ADMGs, all mentioned graphical concepts for DAGs and BGs apply, and define $\text{mb}_\mathcal{G}(V) = \text{pa}_\mathcal{G}(\text{dis}_\mathcal{G}(V)) \cup (\text{dis}_\mathcal{G}(V) \setminus \{V\})$ as the Markov blanket of V in $\mathcal{G}$. Furthermore, a subset $\mathbf{A}$ is called ancestral if whenever $V \in \text{ang}_\mathcal{G}(\mathbf{A})$ then $V \in \mathbf{A}$.

Given a graph $\mathcal{G}$ with a vertex set $\mathbf{V}$, and $\mathbf{A} \subseteq \mathbf{V}$, define $\mathcal{G}_\mathbf{A}$, the *subgraph of $\mathcal{G}$ induced by $\mathbf{A}$* , to be the subgraph of $\mathcal{G}$ containing only the vertex set $\mathbf{A}$ and edges among $\mathbf{A}$.

The list of CIs corresponding to the local Markov property for UGs, DAGs, CGs, BGs and ADMGs with the vertex set $\mathbf{V}$ are as follows

- UGs: $(V \perp\!\!\!\perp \mathbf{V} \setminus (\{V\} \cup \text{ne}_\mathcal{G}(V)) \mid \text{ne}_\mathcal{G}(V))$, for all $V \in \mathbf{V}$.
- DAGs: $(V \perp\!\!\!\perp \text{nd}_\mathcal{G}(V) \setminus \text{pa}_\mathcal{G}(V) \mid \text{pa}_\mathcal{G}(V))$, for all $V \in \mathbf{V}$.
- CGs: $(V \perp\!\!\!\perp \text{nd}_\mathcal{G}(V) \setminus \text{bd}_\mathcal{G}(V) \mid \text{bd}_\mathcal{G}(V))$, for all $V \in \mathbf{V}$.
- BDs: $(V \perp\!\!\!\perp \mathbf{A} \setminus \text{dis}_{\mathcal{G}_\mathbf{A}}(V) \mid \text{dis}_{\mathcal{G}_\mathbf{A}}(V))$, for all V and all subsets $\mathbf{A}$ of $\mathbf{V}$ such that $V \in \mathbf{A}$.
- ADMGs: $(V \perp\!\!\!\perp \mathbf{A} \setminus (\text{mb}_{\mathcal{G}_\mathbf{A}}(V) \cup \{V\}) \mid \text{mb}_{\mathcal{G}_\mathbf{A}}(V))$, for

all V and ancestral set $\mathbf{A}$ such that $V \in \mathbf{A} \subseteq \text{nd}_\mathcal{G}(V)$.

Note that, as expected, the local properties for the UG and DAG model are special cases of the local property for the CG model, and the local properties for the BG and DAG model are special cases of the local property for the ADMG model.

For example, the following are the lists of CIs for the models associated with the graphs in Figure 1.

- The two pathways DAG model:
 $C \perp\!\!\!\perp B \mid A$ and $D \perp\!\!\!\perp A \mid B, C$.
- The undirected square UG model:
 $C \perp\!\!\!\perp B \mid A, D$ and $D \perp\!\!\!\perp A \mid B, C$.
- The dyadic minimal complex CG model:
 $A \perp\!\!\!\perp B, C \perp\!\!\!\perp B \mid A, D$ and $D \perp\!\!\!\perp A \mid B, C$.
- The Bell scenario ADMG model:
 $A \perp\!\!\!\perp B, D$ and $B \perp\!\!\!\perp A, C$.
- The bidirected square scenario ADMG model:
 $A \perp\!\!\!\perp D$ and $B \perp\!\!\!\perp C$.
- The complete DAG: the empty list, as the model is saturated.

Note that there are graphical models associated with ADMGs defined not only in terms of CIs but also generalized CIs, sometimes called “Verma constraints” [Richardson et al., 2023]. As an example, the model associated with Fig. 2 is defined by a CI: $C \perp\!\!\!\perp A \mid B$ and a generalized independence constraint stating that $\sum_B p(D \mid C, B, A)p(B \mid A)$ is a function of only C and D . In general, both CIs and Verma constraints are needed to describe marginal models for hidden variable DAGs. For the purposes of this work, we only consider ADMG models defined exclusively via ordinary CI constraints, called *ordinary Markov models* [Richardson, 2003, Evans, 2018].

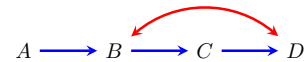


Figure 2: A graph representing a model with a Verma constraint.

2.2 SEMI-PARAMETRIC INFERENCE

As we further describe in Sections 3 and 4, there is a close relationship between Markov model restrictions and the tangent space of the model. This relationship is used to construct estimators via theory of semi-parametric models which we now briefly outline. For further details, we refer the reader to excellent tutorials in Kennedy [2024] and Tsiatis [2006].

We refer to a semi-parametric model as a set of probability densities $\mathcal{P} = \{p(\mathbf{v}; \theta) \mid \theta \in \Theta\}$ relative to a measure μ , parameterized by an infinite dimensional nuisance parameters θ , with the true data generating distribution $p_0 \in \mathcal{P}$.

As an example, the semi-parametric model associated with the DAG model in Fig. 1a is $\mathcal{P}^{(a)} = \{p(a, b, c, d) : p(a, b, c, d) = p(a)p(b | a)p(c | a)p(d | b, c)\}$.

The *contaminated* parametric submodel is a smooth parametric model $\mathcal{P}_\varepsilon = \{p_\varepsilon(\mathbf{v}) | \varepsilon \in \mathcal{E} \subseteq \mathbb{R}\}$ defined around the true distribution p_0 such that $p_{\varepsilon=0} = p_0$, and is contained in the semi-parametric model $\mathcal{P}_\varepsilon \subseteq \mathcal{P}$. In the subsequent discussion, we will refer to $\mathcal{P}_\varepsilon$ as the parametric submodel, following Tsiatis [2006]. We also denote the saturated (unrestricted) model as $\mathcal{P}^{\text{all}}$. As an example, the model associated with the complete DAG in Fig. 1f is saturated. By definition, all semiparametric models are contained in $\mathcal{P}^{\text{all}}$.

For a parametric submodel $\mathcal{P}_\varepsilon$, the score function is the derivative of the log likelihood at the truth $s(\mathbf{v}) = \frac{\partial}{\partial \varepsilon} \log p_\varepsilon(\mathbf{v})|_{\varepsilon=0}$. This function is an element of the Hilbert space $\mathcal{H}$ of all mean-zero $L^2(p_0)$ functions $f(\mathbf{v})$ in which the inner product is relative to the truth distribution $\langle f, g \rangle = \int f(\mathbf{v})g(\mathbf{v})p_0(\mathbf{v})d\mathbf{v}, \forall f, g \in \mathcal{H}$. The tangent space of a semi-parametric model, denoted $\mathcal{T}$, is the closure of all parametric submodels' score functions, thus $\mathcal{T}$ is a subspace of $\mathcal{H}$. For the saturated model $\mathcal{P}^{\text{all}}$, $\mathcal{T} = \mathcal{H}$. The orthogonal complement of the tangent space, denoted $\mathcal{T}^\perp$, is the subspace of all functions in $\mathcal{H}$ orthogonal to $\mathcal{T}$, so $\langle f, g \rangle = 0$ for all $f \in \mathcal{T}^\perp$ and $g \in \mathcal{T}$.

The target of inference in a semi-parametric model is a smooth function $\psi : \mathcal{P} \mapsto \mathbb{R}^d$ where d is a fixed integer. For simplicity, we will only consider scalar target, so $d = 1$. For example, an adjustment formula functional arising in causal inference applications for the examples in Figure 1 yields a scalar parameter $\psi(p) = \mathbb{E}[\mathbb{E}[D | A = a_0, B, C]]$. The main result can be extended easily to $d > 1$ but the notation is more involved.

In semi-parametric theory we are interested in the class of *regular and asymptotically linear (RAL)* estimators for ψ . Given data $\mathbf{V}_1, \dots, \mathbf{V}_n$ drawn from a distribution $p \in \mathcal{P}$, an estimator $\hat{\psi}_n$ for ψ is called asymptotically linear if there is a function $\phi \in \mathcal{H}$ such that

$$\sqrt{n}(\hat{\psi}_n - \psi) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi(\mathbf{V}_i) + o_p(1). \quad (1)$$

By the central limit theorem, the limiting distribution of such an estimator $\hat{\psi}_n$ is normal with root- n rate of convergence,

$$\sqrt{n}(\hat{\psi}_n - \psi) \rightsquigarrow \mathcal{N}(0, \mathbb{E}[\phi^2]), \quad (2)$$

where $\rightsquigarrow$ denotes convergence in distribution. Estimators exhibit asymptotically linear property locally uniformly around the truth are called regular and asymptotically linear (RAL) estimators, which excludes super efficient estimators like the Hodges estimator. The function ϕ is called an *influence function for the estimator $\hat{\psi}_n$* . RAL estimators are attractive because they are root- n consistent and asymp-

totically normal, and the variance equal the variance of its influence function.

A closely related notion is the *influence functions for the target parameter*. Specifically when the distribution changes from p to $\bar{p}$, one can approximate the change in the target functional using the following *von Mises expansion*

$$\psi(\bar{p}) - \psi(p) = \int \varphi(\mathbf{v})(\bar{p} - p)(\mathbf{v})d\mathbf{v} + R[(\bar{p} - p)^2]. \quad (3)$$

The integral is the first-order change of the target parameter with respect to the change in the distribution, while R is the second-order term of the expansion. The function $\varphi \in \mathcal{H}$ characterizes the first-order derivative of the target functional, and is referred to as the influence function of the target parameter ψ . The von Mises expansion implies the following for all parametric submodels p_ε with score function $s \in \mathcal{T}$

$$\left. \frac{\partial}{\partial \varepsilon} \psi(p_\varepsilon) \right|_{\varepsilon=0} = \int \varphi(\mathbf{v})s(\mathbf{v})p_0(\mathbf{v})d\mathbf{v}, \quad (4)$$

where, under regularity conditions, it suffices to consider only parametric submodels of the form $p_\varepsilon(\mathbf{v}) = p_0(\mathbf{v})(1 + \varepsilon s(\mathbf{v}))$. Equation 4 is called the *pathwise differentiability* condition, which is implied by the Neyman orthogonality condition [Chernozhukov et al., 2018], and is central to the derivation of target's influence function. In practice, one first obtains a target parameter's influence function φ from Equation 4, then from it constructs a RAL estimator, whose influence function coincides with the target's influence function, $\phi = \varphi$. We focus on the target parameter's influence functions and will refer to them as just influence functions (IFs).

From Equation 4, it is evident that if φ is an influence function and $f \in \mathcal{T}^\perp$, then $\varphi + f$ is also an influence function, since $\langle f, s \rangle = \int f(\mathbf{v})s(\mathbf{v})p_0(\mathbf{v})d\mathbf{v} = 0$. Therefore, the class of all influence functions is $\varphi \oplus \mathcal{T}^\perp$, where φ is any solution of Equation 4 and $\oplus$ denotes the direct sum (Tsiatis [2006], Theorem 4.3). Thus, knowing the tangent space's orthogonal complement $\mathcal{T}^\perp$ allows one to derive the class of influence functions for any target of inference.

3 THE ORTHOCOMPLEMENT OF THE TANGENT SPACE OF A DAG MODEL

The tangent space of a DAG model and its orthogonal complement have been derived in multiple contexts, see Theorem 4.5 in Tsiatis [2006], Lemma 24 in Rotnitzky and Smucler [2020], and Lemma 3 in Bhattacharya et al. [2022]. In the interests of being self-contained, we repeat the derivation using examples in Figure 1a in order to highlight the main techniques. Despite using specific examples, the derivation is without loss of generality. Subsequently, we explain the reason why these techniques fail to apply to Markov models beyond DAG models.

3.1 A WORKED EXAMPLE

Consider the DAG in Figure 1a, whose semi-parametric model is $\mathcal{P}^{(a)}$. To find the tangent space $\mathcal{T}$, we first derive the score function for an arbitrary parametric submodel p_ε , then take the closure. It turns out that $\mathcal{T}$ is the range of a projection operator Π , hence the orthogonal complement of $\mathcal{T}$ is simply $h - \Pi(h)$ for all $h \in \mathcal{H}$. These steps are demonstrated in detail below.

The score function of a parametric submodel. Consider a parametric submodel $\mathcal{P}_\varepsilon$. By the DAG factorization property, elements of $\mathcal{P}_\varepsilon$ satisfy

$$p_\varepsilon(a, b, c, d) = p_\varepsilon(a)p_\varepsilon(b | a)p_\varepsilon(c | a)p_\varepsilon(d | c, b). \quad (5)$$

This implies that the total score function is the sum

$$s(a, b, c, d) = s(a) + s(b | a) + s(c | a) + s(d | c, b), \quad (6)$$

where, for any pair of disjoint subsets $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$, a component score function $s(\mathbf{x} | \mathbf{y})$ is defined as $\frac{\partial \log p_\varepsilon(\mathbf{x} | \mathbf{y})}{\partial \varepsilon} \Big|_{\varepsilon=0}$.

Since a component score function $s(\mathbf{x} | \mathbf{y})$ has the property that its conditional mean is zero $\int s(\mathbf{x} | \mathbf{y}) p_0(\mathbf{x} | \mathbf{y}) d\mathbf{x} = 0$, the component scores are in the following subspaces of $\mathcal{H}$

$$\begin{aligned} s(a) &\in \mathcal{T}_A \equiv \{\mathbb{E}[h|a] : \forall h \in \mathcal{H}\}, \\ s(b | a) &\in \mathcal{T}_{B|A} \equiv \{\mathbb{E}[h|b, a] - \mathbb{E}[h|a] : \forall h \in \mathcal{H}\}, \\ s(c | a) &\in \mathcal{T}_{C|A} \equiv \{\mathbb{E}[h|c, a] - \mathbb{E}[h|a] : \forall h \in \mathcal{H}\}, \\ s(d | c, b) &\in \mathcal{T}_{D|CB} \equiv \{\mathbb{E}[h|d, c, b] - \mathbb{E}[h|c, b] : \forall h \in \mathcal{H}\}. \end{aligned} \quad (7)$$

The expectations are relative to p_0 . In the Appendix, we show that these subspaces are pairwise orthogonal by checking that the inner products of their elements vanish.

The model's tangent space. The semi-parametric tangent space $\mathcal{T}$ is the closure of all parametric submodels's score functions, hence it must be a closed subspace of the direct sum $\mathcal{T}_A \oplus \mathcal{T}_{B|A} \oplus \mathcal{T}_{C|A} \oplus \mathcal{T}_{D|BC}$. In fact, it is actually this direct sum

$$\mathcal{T} = \mathcal{T}_A \oplus \mathcal{T}_{B|A} \oplus \mathcal{T}_{C|A} \oplus \mathcal{T}_{D|BC}. \quad (8)$$

We refer to the subspaces on the r.h.s as tangent subspaces. Any element of $\mathcal{T}$ can be explicitly constructed as the sum of four elements of the four tangent subspaces, whose explicit forms are given by Equation 7. A detailed proof of Equation 8 is provided in the Appendix. In essence, the equality is established once it is shown that each tangent subspace is contained in $\mathcal{T}$. To illustrate, consider showing that $\mathcal{T}_{B|A} \subseteq \mathcal{T}$. Pick any $f \in \mathcal{T}_{B|A}$ and define the parametric model $p_\varepsilon = p_0(1 + \varepsilon f)$, where ε is small enough so that $p_\varepsilon \geq 0$. The score function at p_0 is $\frac{\partial}{\partial \varepsilon} \log p_\varepsilon \Big|_{\varepsilon=0} = f$. This parametric model contains the true distribution p_0 at $\varepsilon = 0$, and it satisfies all the DAG constraints: $p_\varepsilon(c | b, a) = p_\varepsilon(c | a)$ and $p_\varepsilon(d | c, b, a) = p_\varepsilon(d | c, b)$. Therefore, p_ε defines a valid parametric submodel of $\mathcal{P}$, confirming $\mathcal{T}_{B|A} \subseteq \mathcal{T}$.

The orthogonal complement of the model tangent space. If $\mathcal{S}$ is a closed subspace of $\mathcal{H}$, the *orthogonal complement* of $\mathcal{S}$ in $\mathcal{H}$, denoted $\mathcal{S}^\perp$, is defined as the set of all residues $h - \Pi(h | \mathcal{S})$ for all $h \in \mathcal{H}$, where $\Pi(\cdot | \mathcal{S})$ denotes the projection operator onto the subspace $\mathcal{S}$. Therefore, we first derive $\Pi(\cdot | \mathcal{T})$ before constructing $\mathcal{T}^\perp$.

The expressions of $\mathcal{T}_A, \mathcal{T}_{B|A}, \mathcal{T}_{C|A}, \mathcal{T}_{D|CB}$ in Equation 7 show that these subspaces are ranges of corresponding linear operators $\Lambda_A, \Lambda_{B|A}, \Lambda_{C|A}, \Lambda_{D|CB}$ appeared in these expressions, respectively. For example, the linear operator $\Lambda_{C|A}$ is defined by $\Lambda_{C|A}(h)(c, a) \equiv \mathbb{E}[h | c, a] - \mathbb{E}[h | a]$, and it maps the Hilbert space $\mathcal{H}$ into $\mathcal{T}_{C|A}$. In fact, these operators are actually projection operators onto these subspaces, which can be verified directly by checking the inner product between the residue and the image. For example, in the case of $\mathcal{T}_{C|A}$, we have $\langle h - \Lambda_{C|A}(h), \Lambda_{C|A}(h) \rangle = \mathbb{E}[(h - \mathbb{E}[h | c, a] + \mathbb{E}[h | a])(\mathbb{E}[h | c, a] - \mathbb{E}[h | a])] = 0$ for all $h \in \mathcal{H}$. Since the subspaces $\mathcal{T}_A, \mathcal{T}_{B|A}, \mathcal{T}_{C|A}, \mathcal{T}_{D|CB}$ are pairwise orthogonal, the projection of a function $h \in \mathcal{H}$ onto the tangent space $\mathcal{T}$ is given by the sum of the individual projections

$$\begin{aligned} \Pi(h | \mathcal{T}) &= \Lambda_A(h) + \Lambda_{B|A}(h) + \Lambda_{C|A}(h) + \Lambda_{D|CB}(h) \\ &= \mathbb{E}[h | d, c, b] - \mathbb{E}[h | c, b] + \mathbb{E}[h | c, a] \\ &\quad + \mathbb{E}[h | b, a] - \mathbb{E}[h | a]. \end{aligned} \quad (9)$$

Indeed, one can directly verify that this is the expression of the projection operator onto $\mathcal{T}$ by checking the inner product $\langle h - \Pi(h | \mathcal{T}), \Pi(h | \mathcal{T}) \rangle = 0$ for all $h \in \mathcal{H}$. Hence, the orthogonal complement of the tangent space is

$$\begin{aligned} \mathcal{T}^\perp &= \{h - \mathbb{E}[h | d, c, b] + \mathbb{E}[h | c, b] - \mathbb{E}[h | c, a] \\ &\quad - \mathbb{E}[h | b, a] + \mathbb{E}[h | a] : \forall h \in \mathcal{H}\}. \end{aligned} \quad (10)$$

Given any DAG model, these described steps can be used to derive the tangent space and its orthogonal complement, as done in Tsiatis [2006], Rotnitzky and Smucler [2020], Bhattacharya et al. [2022]. In the special case when the DAG model has only one constraint, we reproduce the following well-known lemma.

Lemma 1. Suppose $\mathcal{P}$ is a single independence DAG model over variables $\mathbf{V}$, defined by a single constraint $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$. The model's tangent space $\mathcal{T}$ and its orthogonal complement $\mathcal{T}^\perp$ are, respectively

$$\begin{aligned} \mathcal{T} &= \{h - \mathbb{E}[h | \mathbf{x}, \mathbf{y}, \mathbf{z}] + \mathbb{E}[h | \mathbf{x}, \mathbf{z}] - \mathbb{E}[h | \mathbf{z}] \\ &\quad + \mathbb{E}[h | \mathbf{y}, \mathbf{z}] : \forall h \in \mathcal{H}\}, \\ \mathcal{T}^\perp &= \{\mathbb{E}[h | \mathbf{x}, \mathbf{y}, \mathbf{z}] - \mathbb{E}[h | \mathbf{y}, \mathbf{z}] \\ &\quad - \mathbb{E}[h | \mathbf{x}, \mathbf{z}] + \mathbb{E}[h | \mathbf{z}] : \forall h \in \mathcal{H}\}. \end{aligned} \quad (11)$$

The proof of this lemma, which is in the Appendix, is similar to the above worked example, except that the model has only one constraint instead of two constraints, and disjoint subsets of variables replace the univariate variables.

3.2 METHODOLOGICAL OBSTACLES FOR OBTAINING GENERAL TANGENT SPACES

In the worked example in the previous section, and indeed in all DAG models, both the tangent space $\mathcal{T}$ and its orthogonal complement $\mathcal{T}^\perp$ are expressible in closed form. That is, one can construct arbitrary elements of $\mathcal{T}$ and $\mathcal{T}^\perp$ using Equation 8 and 10, respectively. This is particularly useful for inference, since if φ is a known influence function for some target parameter $\psi(p)$, then

$$\begin{aligned} \varphi + h - \mathbb{E}[h \mid d, c, b] + \mathbb{E}[h \mid c, b] - \mathbb{E}[h \mid c, a] \\ - \mathbb{E}[h \mid b, a] + \mathbb{E}[h \mid a] \end{aligned} \quad (12)$$

is also an influence function of that target, where h is an arbitrary function in $\mathcal{H}$. The key reason for the existence of the closed form expressions for $\mathcal{T}$ and $\mathcal{T}^\perp$ is the fact that the DAG model likelihood in Equation 5 is also in closed form, as a single factorization representing all constraints in the local Markov property of the model.

On the other hand, in Markov models associated with ADMGs, no single factorization representing all model constraints is known for general state spaces. For example, in the Bell scenario in Figure 1d, a closed form factorization of $p_\varepsilon(a, b, c, d)$ representing both constraints $A \perp\!\!\!\perp B, D$ and $B \perp\!\!\!\perp A, C$ is unknown in general settings. Instead, distributions satisfying both constraints may be obtained iteratively by solving the following system of two equations

$$p_\varepsilon(a, b, c, d) = p_\varepsilon(a)p_\varepsilon(b, d)p_\varepsilon(c \mid d, b, a), \quad (13)$$

$$p_\varepsilon(a, b, c, d) = p_\varepsilon(b)p_\varepsilon(c, a)p_\varepsilon(d \mid c, b, a). \quad (14)$$

The first equation forces the distribution to satisfy $A \perp\!\!\!\perp B, D$, while the second equation forces the likelihood to satisfy $B \perp\!\!\!\perp A, C$. We will refer to them as the likelihood constraint equations. Mathematical objects defined as solutions of some system of equations, such as p_ε above, is said to have *implicit expressions*. Subsequently, the score function of this parametric submodel is also defined implicitly as solutions to the following two equations

$$s(a, b, c, d) = s(a) + s(b, d) + s(c \mid a, b, d), \quad (15)$$

$$s(a, b, c, d) = s(b) + s(a, c) + s(d \mid b, a, c), \quad (16)$$

which are referred to as the score constraint equations. To see that these two equations solve for the same object, we can equivalently reexpress these component score functions in terms of the total score function $s(a, b, c, d)$ as

$$\mathbb{E}[s \mid a, b, d] = \mathbb{E}[s \mid a] + \mathbb{E}[s \mid b, d], \quad (17)$$

$$\mathbb{E}[s \mid b, a, c] = \mathbb{E}[s \mid b] + \mathbb{E}[s \mid a, c]. \quad (18)$$

Similarly to Equations 13 and 14, solving this system of equations is an open problem in general settings. These problems are closely related to implicitization problems in

algebraic geometry [Cox et al., 2015] and parameterization problems of graphical model likelihoods [Evans and Richardson, 2019, Shpitser, 2023], which may be viewed as the problem of transforming an implicit definition of a likelihood surface into an explicit definition of said likelihood indexed by variationally independent parameters.

Absent such a likelihood definition, tangent spaces for graphical models such as the Bell scenario model will have an implicit expression of the form: $\mathcal{T} = \{\text{all common solutions of Equations 17 and 18}\}$. It is unknown how to construct the projection operator onto this subspace. Consequently, the orthogonal complement for the tangent space also has an implicit form: $\mathcal{T}^\perp = \{\text{all } f \text{ orthogonal to all common solutions of 17 and 18}\}$. In other words, as opposed to the DAG case, elements of $\mathcal{T}^\perp$ cannot be written down in closed form.

In the next section, we illustrate an alternative simple approach for obtaining the characterization of the orthogonal complement of the tangent space that bypasses the need for an explicit representation of the likelihood or the tangent space.

4 THE ORTHOCOMPLEMENT OF THE TANGENT SPACE OF A CI MODEL

A general Markov model may be viewed as an intersection of single independence models, each with one CI constraint. Thus, the tangent space of the general Markov model is the intersection of all the single independence models's tangent spaces. Consequently, the orthogonal complement of this intersection is the span of all the orthogonal complements of the single independence models's tangent spaces. This geometric characterization of the tangent space's orthogonal complement for intersection models first appears in Lemma 1.7 of Van Der Laan and Robins [2003]; our argument is an application of this observation in the context of Markov models.

4.1 A WORKED EXAMPLE: THE BELL SCENARIO MODEL

We define the following two models with only one constraint

$$\mathcal{P}_1 = \{\text{all } p(a, b, c, d) \text{ satisfying } A \perp\!\!\!\perp B, D\}, \quad (19)$$

$$\mathcal{P}_2 = \{\text{all } p(a, b, c, d) \text{ satisfying } B \perp\!\!\!\perp A, C\}. \quad (20)$$

The Bell scenario model satisfying both constraints is then the intersection of these two models,

$$\mathcal{P}^{(d)} = \mathcal{P}_1 \cap \mathcal{P}_2. \quad (21)$$

Graphical models are commonly defined in this way, that is as intersections of simpler models [Lauritzen, 1996,

Richardson, 2003, Richardson et al., 2023].

Both models $\mathcal{P}_1$ and $\mathcal{P}_2$ are *single independence DAG models* – DAG models defined by a single constraint. Consequently, the technique described in Section 3 applies to these models. The parametric submodel likelihoods for $\mathcal{P}_1$ and $\mathcal{P}_2$ are Equations 13 and 14, respectively. The score functions for these parametric submodels are given by Equations 15 and 16 (equivalently, Equations 17 and 18), respectively. By Lemma 1, the tangent spaces for these models are

$$\begin{aligned}\mathcal{T}_1 &= \{h - \mathbb{E}[h|a, b, d] + \mathbb{E}[h|a] + \mathbb{E}[h|b, d] : \forall h \in \mathcal{H}\}, \\ \mathcal{T}_2 &= \{h - \mathbb{E}[h|b, a, c] + \mathbb{E}[h|b] + \mathbb{E}[h|a, c] : \forall h \in \mathcal{H}\},\end{aligned}\quad (22)$$

and their orthogonal complements are

$$\begin{aligned}\mathcal{T}_1^\perp &= \{\mathbb{E}[h|a, b, d] - \mathbb{E}[h|a] - \mathbb{E}[h|b, d] : \forall h \in \mathcal{H}\}, \\ \mathcal{T}_2^\perp &= \{\mathbb{E}[h|b, a, c] - \mathbb{E}[h|b] - \mathbb{E}[h|a, c] : \forall h \in \mathcal{H}\}.\end{aligned}\quad (23)$$

The tangent space $\mathcal{T}$ of the Bell scenario model is the closure of all functions satisfying both Equations 17 and 18, $\mathcal{T} = \overline{\mathcal{T}_1 \cap \mathcal{T}_2}$. Since both $\mathcal{T}_1$ and $\mathcal{T}_2$ are closed, $\mathcal{T} = \mathcal{T}_1 \cap \mathcal{T}_2$, the intersection model's tangent space is the intersection of tangent spaces. For detail, see the proof of Theorem 3 in the Appendix. This implies ([Van Der Laan and Robins, 2003], Lemma 1.7)

Lemma 2. *The orthogonal complement of the tangent space of the Bell scenario model is the direct sum $\mathcal{T}^\perp = \mathcal{T}_1^\perp \oplus \mathcal{T}_2^\perp$, whose explicit expression is*

$$\begin{aligned}\mathcal{T}^\perp &= \{\mathbb{E}[h|a, d, b] - \mathbb{E}[h|a] - \mathbb{E}[h|d, b] \\ &\quad + \mathbb{E}[g|b, c, a] - \mathbb{E}[g|b] - \mathbb{E}[g|c, a] : \forall g, h \in \mathcal{H}\}.\end{aligned}\quad (24)$$

Proof. Since $\mathcal{T} = \mathcal{T}_1 \cap \mathcal{T}_2$, elements of $\mathcal{T}$ must be orthogonal to both $\mathcal{T}_1^\perp$ and $\mathcal{T}_2^\perp$. This implies the span $\overline{\mathcal{T}_1^\perp \oplus \mathcal{T}_2^\perp}$ is in the orthogonal complement of $\mathcal{T}$. Since both $\mathcal{T}_1^\perp$ and $\mathcal{T}_2^\perp$ are closed, $\overline{\mathcal{T}_1^\perp \oplus \mathcal{T}_2^\perp} = \mathcal{T}_1^\perp \oplus \mathcal{T}_2^\perp \subseteq \mathcal{T}^\perp$. For other direction, pick any $f \in (\mathcal{T}_1^\perp \oplus \mathcal{T}_2^\perp)^\perp$. Then f must be orthogonal to just $\mathcal{T}_1^\perp$, so $f \in \mathcal{T}_1$. Similarly, $f \in \mathcal{T}_2$, and therefore $f \in \mathcal{T}$. We just show $(\mathcal{T}_1^\perp \oplus \mathcal{T}_2^\perp)^\perp \subseteq \mathcal{T}$, and by taking the orthogonal complement we get $\mathcal{T}^\perp \subseteq \mathcal{T}_1^\perp \oplus \mathcal{T}_2^\perp$. Therefore, $\mathcal{T}^\perp = \mathcal{T}_1^\perp \oplus \mathcal{T}_2^\perp$. Finally, the explicit expression for $\mathcal{T}^\perp$ follows from the definition of direct sum and the explicit expressions for $\mathcal{T}_1^\perp$ and $\mathcal{T}_2^\perp$. $\square$

Consequently, if φ is a known influence function of a target parameter in this model, then

$$\begin{aligned}\varphi &+ \mathbb{E}[h | a, b, d] - \mathbb{E}[h | a] - \mathbb{E}[h | b, d] \\ &+ \mathbb{E}[g | b, a, c] - \mathbb{E}[g | b] - \mathbb{E}[g | a, c]\end{aligned}\quad (25)$$

is also an influence function of that target parameter, where g, h are arbitrary functions in $\mathcal{H}$. Equation 25 thus characterizes the class of influence functions for any target parameter in the Bell scenario model.

4.2 MAIN RESULT

The result for the Bell scenario model described in the previous section can easily be extended to any Markov model, including all mentioned graphical models associated to UGs, CGs, and ADMGs. Consider a general Markov model defined by a list of $K \geq 1$ marginal and conditional independence constraints, as model without any constraint ($K = 0$) is the saturated model described in Tsiatis [2006]. By definition, this Markov model is the intersection of K single independence DAG models defined by a single constraint $\mathcal{P} = \mathcal{P}_1 \cap \dots \cap \mathcal{P}_K$. The tangent space is the intersection of the tangent spaces of these single independence models, $\mathcal{T} = \mathcal{T}_1 \cap \dots \cap \mathcal{T}_K$, and its orthogonal complement is the direct sum $\mathcal{T}^\perp = \mathcal{T}_1^\perp \oplus \dots \oplus \mathcal{T}_K^\perp$.

Theorem 3. *Suppose $\mathcal{P}$ is a general semi-parametric Markov model defined by $K \geq 1$ conditional independence constraints $\mathbf{X}_i \perp\!\!\!\perp \mathbf{Y}_i \mid \mathbf{Z}_i$, where $i = 1, \dots, K$. For each i , let $\mathcal{P}_i$ be the single independence DAG model defined by the i -th constraint. The orthogonal projection operator $\Pi(\cdot \mid \mathcal{T}_i^\perp) : \mathcal{H} \mapsto \mathcal{H}$ sends each function $h \in \mathcal{H}$ to the orthogonal complement $\mathcal{T}_i^\perp$ of the tangent space $\mathcal{T}_i$ of $\mathcal{P}_i$,*

$$\begin{aligned}\Pi(h \mid \mathcal{T}_i^\perp)(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i) &= \mathbb{E}[h|\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i] - \mathbb{E}[h|\mathbf{x}_i, \mathbf{z}_i] \\ &\quad - \mathbb{E}[h|\mathbf{y}_i, \mathbf{z}_i] + \mathbb{E}[h|\mathbf{z}_i].\end{aligned}\quad (26)$$

The orthogonal complement of the tangent space of $\mathcal{P}$ is the direct sum of these orthogonal complements,

$$\mathcal{T}^\perp = \left\{ \sum_{i=1}^K \Pi(h_i \mid \mathcal{T}_i^\perp) : \forall h_1, \dots, h_K \in \mathcal{H} \right\}. \quad (27)$$

Proof Sketch. Applying Lemma 1 to the single independence model $\mathcal{P}_i$, any element of the tangent space's orthogonal complement $\mathcal{T}_i^\perp$ is $\Pi(h_i \mid \mathcal{T}_i^\perp)$, with arbitrary $h_i \in \mathcal{H}$. A general version of Lemma 2 shows that $\mathcal{T}^\perp = \mathcal{T}_1^\perp \oplus \dots \oplus \mathcal{T}_K^\perp$. Hence, any element of $\mathcal{T}^\perp$ is a sum $\sum_{i=1}^K \Pi(h_i \mid \mathcal{T}_i^\perp)$ of elements of $\{\mathcal{T}_i^\perp\}_{i=1}^K$, where $\{h_i \in \mathcal{H}\}_{i=1}^K$ are arbitrary (not necessarily equal) functions. $\square$

This theorem applies to all Markov models – models defined by a list of CI constraints – which are not necessarily graphical models. Furthermore, for graphical Markov models such as those associated with UGs, DAGs, CGs, BGs, and ADMGs, the list of CI constraints can be obtained from global, local or any other Markov property, as long as this list fully describes the model. This means that it is possible to obtain several syntactically different representations of the same orthocomplement of the tangent space of the model by using different Markov properties. For example, the local Markov property for DAGs listed in Section 2.1 differs from the ordered local Markov property for DAGs, which is used in Bhattacharya et al. [2022], Tsiatis [2006],

Rotnitzky and Smucler [2020]. Despite the different expressions, the orthogonal complement of the tangent space is the same.

Equation 79 suggests an operator $\Gamma : \mathcal{H} \mapsto \mathcal{T}^\perp$ sending functions in $\mathcal{H}$ to functions in the orthogonal complement $\mathcal{T}^\perp$, defined by $\Gamma(f) = \sum_{i=1}^K \Pi(f | \mathcal{T}_i^\perp)$. For general Markov model, it is unknown whether this operator is the projection operator onto $\mathcal{T}^\perp$. The Appendix shows an example in the Bell scenario case. Therefore, finding the projection operator onto a general Markov model's tangent space or its orthogonal complement is still an open problem. Since the efficient influence function (EIF) is $\varphi - \Pi(\varphi | \mathcal{T}^\perp)$, where φ is any influence function and $\Pi(\cdot | \mathcal{T}^\perp)$ is the projection onto the orthogonal complement $\mathcal{T}^\perp$ of the tangent space, the exact form of the efficient influence function for a target parameter in a general Markov model is also an open problem.

A direct consequence of Theorem 3 is the explicit characterization of the class of all influence functions of any target parameter $\psi(p)$ in a general Markov model. If φ is a known influence function of $\psi(p)$, then

$$\varphi + \sum_{i=1}^K \Pi(h_i | \mathcal{T}_i^\perp) \quad (28)$$

is also an influence function of $\psi(p)$, for any $\{h_i \in \mathcal{H}\}_{i=1}^K$. An explicit example is shown in the next section.

5 APPLICATIONS

5.1 THE CLASS OF INFLUENCE FUNCTIONS

In this subsection, we illustrate how our result streamline the derivations of the class of influence functions for the conditional mean parameter $\psi(p) = \mathbb{E}[D | A = a_0]$ with a fixed value a_0 , for all the Markov models corresponding to the graphs in Figure 1. Extensions to other target parameters are trivial. For example, if the target parameter is the adjustment formula $\mathbb{E}[\mathbb{E}[D | A = a_0, C]]$, simply replace φ below by the augmented inverse probability weighting (AIPW) functional.

The saturated model $\mathcal{P}_{ABCD}^{\text{all}}$ in Figure 1f. The saturated model $\mathcal{P}_{ABCD}^{\text{all}}$ over four variables $\{A, B, C, D\}$ is a saturated extension over the saturated model $\mathcal{P}_{AD}^{\text{all}}$ with only two variables $\{A, D\}$, with B and C play the role of auxiliary variables. Following Tsiatis [2006], Chapter 5.5, the set of influence functions in $\mathcal{P}_{ABCD}^{\text{all}}$ is exactly the set of influence functions in $\mathcal{P}_{AD}^{\text{all}}$. This means, for the purpose of estimating $\mathbb{E}[D | A = a_0]$, we can safely ignore B and C as these variables do not provide efficiency gain. Moreover, the derivation of the influence functions in $\mathcal{P}_{AD}^{\text{all}}$ involves only two variables and hence is much simpler.

Any saturated model has $\{0\}$ as its orthogonal complement of the tangent space, so there is only one influence function,

$$\varphi(a, b, c, d) = \frac{\mathbb{I}(a = a_0)}{p(a)}(d - \mathbb{E}[D|a]). \quad (29)$$

All other Markov models corresponding to the graphs in Figure 1a-e are submodels of $\mathcal{P}_{ABCD}^{\text{all}}$. Therefore, φ is also an influence function in any of these models. The other ingredient for the derivation the class of influence functions in these models is the orthogonal complement of the tangent space given by Theorem 3.

The DAG model $\mathcal{P}^{(a)}$ in Figure 1a. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be the simple models corresponding to the constraints $C \perp\!\!\!\perp B | A$ and $D \perp\!\!\!\perp A | C, B$, respectively. Let $\mathcal{T}_1^\perp$ and $\mathcal{T}_2^\perp$ denote the orthocomplements of their respective tangent spaces. The projection operators onto these orthocomplements are

$$\begin{aligned} \Pi_a(g | \mathcal{T}_1^\perp) &= \mathbb{E}[g|c, b, a] - \mathbb{E}[g|c, a] - \mathbb{E}[g|b, a] + \mathbb{E}[g|a], \\ \Pi_a(h | \mathcal{T}_2^\perp) &= h - \mathbb{E}[h|d, c, b] - \mathbb{E}[h|a, c, b] + \mathbb{E}[h|c, b]. \end{aligned}$$

Therefore, the orthogonal complement of the tangent space $\mathcal{T}$ for this DAG model is, according to Equation 79

$$\mathcal{T}^\perp = \{\Pi_a(g | \mathcal{T}_1^\perp) + \Pi_a(h | \mathcal{T}_2^\perp) : \forall g, h \in \mathcal{H}\}, \quad (30)$$

This expression differs from the one given in Equation 10. To reconcile the two, note that $\mathcal{T}^\perp$ is the direct sum of the orthogonal subspaces $\mathcal{T}_1^\perp$ and $\mathcal{T}_2^\perp$. This geometric property holds for general DAGs: the orthocomplements of the tangent spaces corresponding to distinct CIs in the local Markov property are pairwise orthogonal. Since the projection operators onto these subspaces are known – namely, $\Pi_a(\cdot | \mathcal{T}_1^\perp)$ and $\Pi_a(\cdot | \mathcal{T}_2^\perp)$ – the projection of a function $h \in \mathcal{H}$ onto $\mathcal{T}^\perp$ is given by $\Pi(h | \mathcal{T}^\perp) = \Pi_a(h | \mathcal{T}_1^\perp) + \Pi_a(h | \mathcal{T}_2^\perp)$. This shows that the two expressions for $\mathcal{T}^\perp$ in 10 and 30 describe the same subspace.

The class of influence functions for $\psi(p)$ in this model is $\{\varphi + \Pi_a(g | \mathcal{T}_1^\perp) + \Pi_a(h | \mathcal{T}_2^\perp) : \forall g, h \in \mathcal{H}\}$.

The UG model $\mathcal{P}^{(b)}$ in Figure 1b. $\mathcal{P}_1$ and $\mathcal{P}_2$ correspond to $C \perp\!\!\!\perp B | A, D$ and $D \perp\!\!\!\perp A | C, B$, respectively.

$$\begin{aligned} \Pi_b(g | \mathcal{T}_1^\perp) &= g - \mathbb{E}[g | c, a, d] - \mathbb{E}[g|b, a, d] + \mathbb{E}[g|a, d], \\ \Pi_b(h | \mathcal{T}_2^\perp) &= h - \mathbb{E}[h|d, c, b] - \mathbb{E}[h|a, c, b] + \mathbb{E}[h|c, b]. \end{aligned}$$

The class of influence functions for $\psi(p)$ in this model is $\{\varphi + \Pi_b(g | \mathcal{T}_1^\perp) + \Pi_b(h | \mathcal{T}_2^\perp) : \forall g, h \in \mathcal{H}\}$.

The CG model $\mathcal{P}^{(c)}$ in Figure 1c. $\mathcal{P}_1$ and $\mathcal{P}_2$ are the same models in the UG case, while $\mathcal{P}_3$ corresponds to $A \perp\!\!\!\perp B$. The projection operator onto $\mathcal{T}_3^\perp$ is

$$\Pi_c(f | \mathcal{T}_3^\perp) = \mathbb{E}[f | a, b] - \mathbb{E}[f | a] - \mathbb{E}[f | b].$$

The class of influence functions for $\psi(p)$ in this model is $\{\varphi + \Pi_b(g | \mathcal{T}_1^\perp) + \Pi_b(h | \mathcal{T}_2^\perp) + \Pi_c(f | \mathcal{T}_3^\perp) : \forall f, g, h \in \mathcal{H}\}$.

The bidirected square model $\mathcal{P}^{(e)}$ in Figure 1e. $\mathcal{P}_1$ and $\mathcal{P}_2$ correspond to $A \perp\!\!\!\perp D$ and $B \perp\!\!\!\perp C$, respectively.

$$\begin{aligned}\Pi_d(g \mid \mathcal{T}_1^\perp) &= \mathbb{E}[g \mid a, d] - \mathbb{E}[g \mid a] - \mathbb{E}[g \mid d], \\ \Pi_d(h \mid \mathcal{T}_2^\perp) &= \mathbb{E}[h \mid b, c] - \mathbb{E}[h \mid b] - \mathbb{E}[h \mid c].\end{aligned}$$

The class of influence functions for $\psi(p)$ in this model is $\{\varphi + \Pi_d(g \mid \mathcal{T}_1^\perp) + \Pi_d(h \mid \mathcal{T}_2^\perp) : \forall g, h \in \mathcal{H}\}$.

5.2 IMPROVING EFFICIENCY

The variance of a RAL estimator $\hat{\psi}$ for a target parameter $\psi(p)$ is exactly the variance of the corresponding influence function $\varphi \in \mathcal{H}$. For influence functions have mean zero, the variance of φ is its squared norm $\mathbb{E}[\varphi^2]$. Thus, an estimator $\hat{\psi}_2$ is at least as efficient as another estimator $\hat{\psi}_1$ precisely when the variance of the influence function φ_2 of $\hat{\psi}_2$ is equal or smaller than the variance of the influence functions φ_1 of $\hat{\psi}_1$, $\mathbb{E}[\varphi_2^2] \leq \mathbb{E}[\varphi_1^2]$ (Tsiatis [2006], Chapter 3). The efficient influence function is simply the influence function with the smallest variance in the class of all influence functions of the model. Therefore, in order to construct more efficient estimators, the first step is finding influence functions with smaller variances.

Given an initial influence function φ_0 of a target parameter, our main result immediately implies an iterative method for constructing sequences of influence functions with non-increasing variances. As an example, suppose the Markov model has $K \geq 2$ CI constraints. Theorem 3 states that any $\sum_{i=1}^K \Pi(h_i \mid \mathcal{T}_i^\perp)$ is an element of $\mathcal{T}^\perp$. Therefore, $\Pi(\varphi_0 \mid \mathcal{T}_1^\perp) \in \mathcal{T}^\perp$. By Theorem 4.3 of Tsiatis [2006], adding an element in $\mathcal{T}^\perp$ to an influence function yields another influence function of the same target, hence

$$\varphi_1 := \varphi_0 - \Pi(\varphi_0 \mid \mathcal{T}_1^\perp) \quad (31)$$

is an influence function. Moreover, since $\Pi(\varphi_0 \mid \mathcal{T}_1^\perp)$ is the orthogonal projection of φ_0 onto $\mathcal{T}_1^\perp$, the above φ_1 is the orthogonal projection of φ_0 onto $\mathcal{T}_1$. The Pythagorean theorem for Hilbert spaces (Tsiatis [2006], Theorem 3.3) states that the projection of a vector cannot be longer than the vector itself, so $\mathbb{E}[\varphi_1^2] \leq \mathbb{E}[\varphi_0^2]$. This implies that φ_1 is an influence function with equal or smaller variance than φ_0 . Next, consider

$$\varphi_2 := \varphi_1 - \Pi(\varphi_1 \mid \mathcal{T}_2^\perp). \quad (32)$$

By the same reasoning, φ_2 is an influence function with equal or smaller variance than φ_1 . Repeating this process M times, we obtain a sequence of influence functions $\varphi_0, \varphi_1, \varphi_2, \dots, \varphi_M$, with the property that the variance of φ_m is equal or smaller than the variance of φ_{m-1} , for all $m = 1, \dots, M$. The influence function φ_m is referred to as an m -improvement of the initial φ_0 . If we can construct an estimator $\hat{\psi}_M$ based on the influence function φ_M , then this

estimator is at least as efficient as the estimator $\hat{\psi}_0$ based on the initial influence function φ_0 . In the Appendix, we give a concrete example of such a construction with $M = 2$.

Theorem 4. Consider a Markov model with $K \geq 1$ constraint(s), and let φ_0 be an influence function of a target parameter $\psi(p)$. For any integer sequence $i_1, i_2, \dots, i_M$, where $i_m \in \{1, \dots, K\}$ for all $m \in \{1, \dots, M\}$, define

$$\varphi_m := \varphi_{m-1} - \Pi(\varphi_{m-1} \mid \mathcal{T}_{i_m}^\perp), \quad \forall m \in \{1, \dots, M\}. \quad (33)$$

In words, φ_m is the projection of φ_{m-1} onto the tangent space $\mathcal{T}_{i_m}$ of the DAG model corresponding to the i_m -th constraint. Then $\varphi_0, \dots, \varphi_M$ is a sequence of influence functions of $\psi(p)$ with non-increasing variance, meaning $\mathbb{E}[\varphi_m^2] \leq \mathbb{E}[\varphi_{m-1}^2]$ for all $m \in \{1, \dots, M\}$.

If $i_{m+1} = i_m$, then $\varphi_{m+1} = \varphi_m$ is the same influence function, because $\varphi_m \in \mathcal{T}_{i_m}$ and so $\Pi(\varphi_m \mid \mathcal{T}_{i_{m+1}}^\perp) = 0$. Then, for single constraint models ($K = 1$), there is only one improvement, $\varphi = \varphi_0 - \Pi(\varphi_0 \mid \mathcal{T}^\perp)$, which is exactly the efficient influence function in these models. In models with $K > 1$, the connection between the efficient influence function and φ_M as $M \rightarrow \infty$ is an interesting question, yet outside the scope of our paper.

6 CONCLUSION

We derived an explicit expression for the orthogonal complement of the tangent space for Markov models – statistical models defined using only marginal and conditional independence constraints.

We demonstrated two applications of our result in these models. The first one is an explicit characterization of the class of influence functions for any target parameter. The second one is an iterative method to improve efficiency of any initial RAL estimator of any target parameter.

Our method does not yield an explicit expression for the likelihood or the tangent space of these models, so these questions remain open. Another important open problem is the projection of a function $h \in \mathcal{H}$ onto the tangent space of these models. Derivation of such a projection is equivalent to the derivation of the efficient influence function in a semi-parametric Markov model. Our method provides an important step towards achieving this goal by providing the explicit form of the orthogonal complement of the tangent space.

Acknowledgements

This work was supported by the Office of Naval Research (Grant No. N000142412701), the National Institutes of Health (Grant No. 1R56AI191526-01 and Grant no. 1R01LM014800-01A1), and the National Science Foundation CAREER Award (Grant No. 1942239).

References

- Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric Inference For Causal Effects In Graphical Models With Hidden Variables. *Journal of Machine Learning Research*, 23(295):1–76, 2022. ISSN 1533-7928.
- Peter J. Bickel, Chris A. J. Klaassen, Ya’acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York Berlin Heidelberg, 1993. ISBN 978-0-8018-4541-3.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, February 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097.
- David A. Cox, John Little, and Donal O’Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Undergraduate Texts in Mathematics. Springer International Publishing, Cham, 2015. ISBN 978-3-319-16720-6 978-3-319-16721-3. doi: 10.1007/978-3-319-16721-3.
- Robin J. Evans. Margins of discrete bayesian networks. *Annals of Statistics*, 46:2623–2656, 2018.
- Robin J. Evans and Thomas S. Richardson. Smooth, Identifiable Supermodels of Discrete DAG Models with Latent Variables. *Bernoulli*, 25(2):848–876, May 2019. ISSN 1350-7265. doi: 10.3150/17-BEJ1005.
- M. Frydenberg. The chain graph Markov property. *Scandinavian Journal of Statistics*, 1990.
- Edward H. Kennedy. Semiparametric Doubly Robust Targeted Double Machine Learning: A Review. In *Handbook of Statistical Methods for Precision Medicine*. Chapman and Hall/CRC, 2024.
- Steffen L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Clarendon Press ; Oxford University Press, Oxford, New York, 1996. ISBN 978-0-19-852219-5.
- Steffen L. Lauritzen and Thomas S. Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):321–348, 2002. ISSN 1467-9868. doi: 10.1111/1467-9868.00340.
- H. F. Lopes, E. Salazar, and D. Gamerman. Spatial dynamic factor analysis. *Bayesian Analysis*, 3(4):759–792, 2008.
- Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright, editors. *Handbook of Graphical Models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, November 2018. ISBN 978-0-429-46397-6. doi: 10.1201/9780429463976.
- Whitney K. Newey. Semiparametric Efficiency Bounds. *Journal of Applied Econometrics*, 5(2):99–135, 1990. ISSN 1099-1255. doi: 10.1002/jae.3950050202.
- Elizabeth L. Ogburn, Ilya Shpitser, and Youjin Lee. Causal Inference, Social Networks and Chain Graphs. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(4):1659–1676, October 2020. ISSN 0964-1998. doi: 10.1111/rssa.12594.
- Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, s.l., 1988. ISBN 978-0-08-051489-5. doi: 10.1016/C2009-0-27609-4.
- Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, U.K. ; New York, 2000. ISBN 978-0-521-89560-6.
- Thomas S. Richardson. Markov Properties for Acyclic Directed Mixed Graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003. ISSN 0303-6898.
- Thomas S. Richardson and James M. Robins. Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality, 2013.
- Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1): 334–361, February 2023. ISSN 0090-5364, 2168-8966. doi: 10.1214/22-AOS2253.
- Andrea Rotnitzky and Ezequiel Smucler. Efficient Adjustment Sets for Population Average Causal Treatment Effect Estimation in Graphical Models. *Journal of Machine Learning Research*, 21(188):1–86, 2020. ISSN 1533-7928.
- K. Sachs, O. Perez, D. Pe’er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.
- E. Segal, M. Shapira, A. Regev, D. Pe’er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2): 166–76, 2003.
- Ilya Shpitser. The Lauritzen-Chen Likelihood For Graphical Models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 4181–4195. PMLR, April 2023.

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 - Volume 2*, pages 1219–1226, July 2006. ISBN 978-1-57735-281-5.

Ilya Shpitser, Thomas S. Richardson, and James M. Robins. Multivariate Counterfactual Systems and Causal Graphical Models. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36, pages 813–852. Association for Computing Machinery, New York, NY, USA, 1 edition, March 2022. ISBN 978-1-4503-9586-1.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. A Bradford Book, Cambridge, Mass, second edition edition, January 2001. ISBN 978-0-262-19440-2.

Eric J. Tchetgen Tchetgen, Isabel R. Fulcher, and Ilya Shpitser. Auto-G-Computation of Causal Effects on a Network. *Journal of the American Statistical Association*, 116(534):833–844, April 2021. ISSN 0162-1459. doi: 10.1080/01621459.2020.1811098.

Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer, New York, NY, 2006. ISBN 978-0-387-32448-7. doi: 10.1007/0-387-37345-4.

Mark J. Van Der Laan and James M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer, New York, NY, 2003. ISBN 978-1-4419-3055-2 978-0-387-21700-0. doi: 10.1007/978-0-387-21700-0.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 2000. ISBN 978-0-521-78450-4.

Aad Van Der Vaart. Semiparametric Statistics. In *Lectures on Probability Theory and Statistics*, volume 1781 of *Lecture Notes in Mathematics*, pages 331–457. 2002.

A Characterization of the Orthocomplement of the Tangent Space of Semiparametric Markov Models (Supplementary Material)

Trung Phung¹

Ilya Shpitser¹

¹Computer Science Department, Johns Hopkins University, Baltimore, Maryland, USA

A ADDITIONAL DETAILS ABOUT SEMIPARAMETRIC THEORY

The set of all influence function and the orthogonal complement of the tangent space. As mentioned in the main paper, an influence function of the target parameter $\psi(p)$ is a solution of the pathwise differentiability condition, for all parametric submodel p_ε (Tsiatis [2006], Theorem 3.2 and 4.2)

$$\left. \frac{\partial}{\partial \varepsilon} \psi(p_\varepsilon) \right|_{\varepsilon=0} = \int \varphi(\mathbf{v}) s(\mathbf{v}) p_0(\mathbf{v}) d\mathbf{v}. \quad (4)$$

Suppose φ is an influence function of $\psi(p)$ and $\mathcal{T}^\perp$ is the orthogonal complement of the tangent space of the model. Let f be any function in $\mathcal{T}^\perp$. Since the score function $s(\mathbf{v}) \in \mathcal{T}$, it is orthogonal to f , that is $\langle f, s \rangle = \int f(\mathbf{v}) s(\mathbf{v}) p_0(\mathbf{v}) d\mathbf{v} = 0$. Then $\varphi + f$ is also an influence function, because

$$\left. \frac{\partial}{\partial \varepsilon} \psi(p_\varepsilon) \right|_{\varepsilon=0} = \int [\varphi(\mathbf{v}) + f(\mathbf{v})] s(\mathbf{v}) p_0(\mathbf{v}) d\mathbf{v}. \quad (34)$$

On the other hand, if $\tilde{\varphi}$ is any other influence function, meaning

$$\left. \frac{\partial}{\partial \varepsilon} \psi(p_\varepsilon) \right|_{\varepsilon=0} = \int \tilde{\varphi}(\mathbf{v}) s(\mathbf{v}) p_0(\mathbf{v}) d\mathbf{v}. \quad (35)$$

Then, for all parametric submodel p_ε (equivalently, for all score function $s(\mathbf{v}) \in \mathcal{T}$)

$$0 = \int [\tilde{\varphi}(\mathbf{v}) - \varphi(\mathbf{v})] s(\mathbf{v}) p_0(\mathbf{v}) d\mathbf{v}. \quad (36)$$

This means $\tilde{\varphi} - \varphi \in \mathcal{T}^\perp$, i.e., $\tilde{\varphi}$ is the sum of φ and an element of the orthogonal complement of the tangent space. This result is Theorem 4.3 in Tsiatis [2006], stating that: the set of all influence functions of a target parameter in a semiparametric model is the linear variety $\varphi \oplus \mathcal{T}^\perp := \{\varphi + f : f \in \mathcal{T}^\perp\}$, where φ is any influence function and $\mathcal{T}^\perp$ is the orthogonal complement of the tangent space. Hence, knowing the orthogonal complement of the tangent space $\mathcal{T}^\perp$ is vital to characterize the set of all influence functions.

Variance of an influence function and efficiency of an estimator. As stated in the main paper, the limiting distribution of a RAL estimator $\hat{\psi}_n$ of the target parameter ψ is

$$\sqrt{n}(\hat{\psi}_n - \psi) \rightsquigarrow \mathcal{N}(0, \mathbb{E}[\varphi^2]) \quad (2)$$

where φ is the influence function corresponding to $\hat{\psi}$ [Tsiatis, 2006]. Since $\varphi \in \mathcal{H}$, it has zero mean, and therefore the variance of the influence function φ is its square norm in $\mathcal{H}$, $\text{Var}(\varphi) = E[\varphi^2] - E[\varphi]^2 = E[\varphi^2] = \langle \varphi, \varphi \rangle$.

An estimator $\hat{\psi}$ is more efficient than another estimator $\hat{\psi}'$ precisely when the variance of the limiting distribution of $\sqrt{n}(\hat{\psi}_n - \psi)$ is equal or smaller than the variance of the limiting distribution of $\sqrt{n}(\hat{\psi}'_n - \psi)$ [Tsiatis, 2006]. This means

the variance of the influence function φ of $\hat{\psi}$ is equal or smaller than the variance of the influence function φ' of $\hat{\psi}'$. Therefore, given an influence function φ and the orthogonal complement of the tangent space $\mathcal{T}^\perp$, if one can construct $\varphi - f$, where $f \in \mathcal{T}^\perp$, such that $\mathbb{E}[(\varphi - f)^2] \leq \mathbb{E}[\varphi^2]$, then the estimator obtained from $\varphi - f$ will be equal or more efficient than the estimator obtained from φ . In particular, we derive the orthogonal complement of the tangent space for general Markov model in Theorem 3, and propose a method to improve the variance of an initial influence function in Theorem 4.

The efficient influence function is the influence function $\varphi^{(e)}$ in the class of all influence functions $\varphi \oplus \mathcal{T}^\perp$ with smallest variance, i.e., $\mathbb{E}[(\varphi^{(e)})^2] \leq \mathbb{E}[\varphi'^2]$ for any influence function $\varphi' \in \varphi \oplus \mathcal{T}^\perp$ [Tsiatis, 2006]. One way to find the efficient influence function is projecting any other influence function φ onto the tangent space $\mathcal{T}$. Let $\Pi(\cdot | \mathcal{T})$ be the projection onto $\mathcal{T}$, and $\Pi(\cdot | \mathcal{T}^\perp)$ is the projection onto the orthogonal complement $\mathcal{T}^\perp$. We have $\Pi(\cdot | \mathcal{T}^\perp) = I - \Pi(\cdot | \mathcal{T})$. For any influence function φ

$$\varphi = \varphi - \Pi(\varphi | \mathcal{T}^\perp) + \Pi(\varphi | \mathcal{T}^\perp) \quad (37)$$

Since $\Pi(\varphi | \mathcal{T}^\perp)$ is an element in the orthogonal complement of the tangent space, $\varphi - \Pi(\varphi | \mathcal{T}^\perp)$ must be another influence function. Since $\varphi - \Pi(\varphi | \mathcal{T}^\perp) = \Pi(\varphi | \mathcal{T})$ is the projection of φ onto a subspace, its norm must be equal or less than the norm of φ , so $\mathbb{E}[(\varphi - \Pi(\varphi | \mathcal{T}^\perp))^2] \leq \mathbb{E}[\varphi^2]$. It can be shown (Tsiatis [2006], Theorem 3.5) that the efficient influence function is $\varphi^{(e)} = (\varphi - \Pi(\varphi | \mathcal{T}^\perp))$. Hence, to find the efficient influence function, one must first know the projection operator $\Pi(\cdot | \mathcal{T}^\perp)$ or $\Pi(\cdot | \mathcal{T})$.

B PROOFS IN SECTION 3

The following lemma is the general version of Equation 7.

Lemma 5 (Tsiatis [2006], Theorem 4.5). *Let $\mathcal{P}$ be a semi-parametric model over variables $\mathbf{V}$. Consider any disjoint subsets $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$. For any parametric submodel p_ε , the score function $s(\mathbf{x} | \mathbf{y}) = \frac{\partial}{\partial \varepsilon} \log p_\varepsilon(\mathbf{x} | \mathbf{y})|_{\varepsilon=0}$ at p_0 is an element of the subspace*

$$\mathcal{T}_{\mathbf{X}|\mathbf{Y}} := \{E[h | \mathbf{x}, \mathbf{y}] - E[h | \mathbf{y}] | h \in \mathcal{H}\}. \quad (38)$$

Proof. Let $\mathbf{Z} = \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$. The following is a really useful identity for the score function

$$s(\mathbf{x}, \mathbf{y}, \mathbf{z}) := \frac{\partial}{\partial \varepsilon} \log p_\varepsilon(\mathbf{x}, \mathbf{y}, \mathbf{z}) \Big|_{\varepsilon=0} = \frac{1}{p_0(\mathbf{x}, \mathbf{y}, \mathbf{z})} \frac{\partial}{\partial \varepsilon} p_\varepsilon(\mathbf{x}, \mathbf{y}, \mathbf{z}) \Big|_{\varepsilon=0}. \quad (39)$$

Then

$$\begin{aligned} s(\mathbf{x} | \mathbf{y}) &:= \frac{\partial}{\partial \varepsilon} \log p_\varepsilon(\mathbf{x} | \mathbf{y}) \Big|_{\varepsilon=0} = \frac{\partial}{\partial \varepsilon} \log p_\varepsilon(\mathbf{x}, \mathbf{y}) \Big|_{\varepsilon=0} - \frac{\partial}{\partial \varepsilon} \log p_\varepsilon(\mathbf{y}) \Big|_{\varepsilon=0} \\ &= \frac{1}{p_0(\mathbf{x}, \mathbf{y})} \frac{\partial}{\partial \varepsilon} \int p_\varepsilon(\mathbf{x}, \mathbf{y}, \mathbf{z}) d\mathbf{z} \Big|_{\varepsilon=0} - \frac{1}{p_0(\mathbf{y})} \frac{\partial}{\partial \varepsilon} \int p_\varepsilon(\mathbf{x}, \mathbf{y}, \mathbf{z}) d\mathbf{x} d\mathbf{z} \Big|_{\varepsilon=0} \\ &= \frac{1}{p_0(\mathbf{x}, \mathbf{y})} \int \left(\frac{\partial}{\partial \varepsilon} p_\varepsilon(\mathbf{x}, \mathbf{y}, \mathbf{z}) \Big|_{\varepsilon=0} \right) d\mathbf{z} - \frac{1}{p_0(\mathbf{y})} \int \left(\frac{\partial}{\partial \varepsilon} p_\varepsilon(\mathbf{x}, \mathbf{y}, \mathbf{z}) \Big|_{\varepsilon=0} \right) d\mathbf{x} d\mathbf{z} \\ &= \frac{1}{p_0(\mathbf{x}, \mathbf{y})} \int s(\mathbf{x}, \mathbf{y}, \mathbf{z}) p_0(\mathbf{x}, \mathbf{y}, \mathbf{z}) d\mathbf{z} - \frac{1}{p_0(\mathbf{y})} \int s(\mathbf{x}, \mathbf{y}, \mathbf{z}) p_0(\mathbf{x}, \mathbf{y}, \mathbf{z}) d\mathbf{x} d\mathbf{z} \\ &= \int s(\mathbf{x}, \mathbf{y}, \mathbf{z}) p_0(\mathbf{z} | \mathbf{x}, \mathbf{y}) d\mathbf{z} - \int s(\mathbf{x}, \mathbf{y}, \mathbf{z}) p_0(\mathbf{x}, \mathbf{z} | \mathbf{y}) d\mathbf{x} d\mathbf{z} \\ &= E[s(\mathbf{V}) | \mathbf{x}, \mathbf{y}] - E[s(\mathbf{V}) | \mathbf{y}]. \end{aligned} \quad (40)$$

Since $s(\mathbf{v}) \in \mathcal{H}$, the claim is established. $\square$

The following lemma shows that the tangent subspaces of the two pathway DAG model (Figure 1a) described in Equation 7 are orthogonal.

Lemma 6. *Let $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ be disjoint sets of random variables. The following 2 subspaces are orthogonal*

$$\begin{aligned} \mathcal{T}_1 &= \{E[h | \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}] - E[h | \mathbf{b}, \mathbf{c}, \mathbf{d}] : \forall h \in \mathcal{H}\} \\ \mathcal{T}_2 &= \{E[h | \mathbf{c}, \mathbf{d}] - E[h | \mathbf{d}] : \forall h \in \mathcal{H}\} \end{aligned} \quad (41)$$

Proof. Pick any $\mathbb{E}[h_1 \mid \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}] - \mathbb{E}[h_1 \mid \mathbf{b}, \mathbf{c}, \mathbf{d}] \in \mathcal{T}_1$ and any $\mathbb{E}[h_2 \mid \mathbf{c}, \mathbf{d}] - \mathbb{E}[h_2 \mid \mathbf{d}] \in \mathcal{T}_2$, then consider their inner product

$$\begin{aligned} & \mathbb{E} \left[(\mathbb{E}[h_1 \mid \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}] - \mathbb{E}[h_1 \mid \mathbf{B}, \mathbf{C}, \mathbf{D}]) (\mathbb{E}[h_2 \mid \mathbf{C}, \mathbf{D}] - \mathbb{E}[h_2 \mid \mathbf{D}]) \right] \\ &= \int \left(\int (\mathbb{E}[h_1 \mid \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}] - \mathbb{E}[h_1 \mid \mathbf{b}, \mathbf{c}, \mathbf{d}]) (\mathbb{E}[h_2 \mid \mathbf{c}, \mathbf{d}] - \mathbb{E}[h_2 \mid \mathbf{d}]) p(\mathbf{a}, \mathbf{b} \mid \mathbf{c}, \mathbf{d}) d\mathbf{a} d\mathbf{b} \right) p(\mathbf{c}, \mathbf{d}) d\mathbf{c} d\mathbf{d} \\ &= 0 \end{aligned} \quad (42)$$

□

Lemma 7. *The tangent space of the two pathway DAG model in Figure 1a is the direct sum*

$$\mathcal{T} = \mathcal{T}_A \oplus \mathcal{T}_{B|A} \oplus \mathcal{T}_{C|A} \oplus \mathcal{T}_{D|BC} \quad (8)$$

Proof. Equation 7 shows that $\mathcal{T} \subseteq \mathcal{T}_A \oplus \mathcal{T}_{B|A} \oplus \mathcal{T}_{C|A} \oplus \mathcal{T}_{D|BC}$. To prove the equality, all we need to prove is that each subspace on the r.h.s. is contained in $\mathcal{T}$. In each case, we pick arbitrary f in the subspace under consideration and define a parametric model $\mathcal{P}_\varepsilon$ whose elements are $p_\varepsilon = p_0(1 + \varepsilon f)$. We ensure p_ε is a valid density using the conventional technique in Tsiatis [2006]: (i) non-negativity $p_\varepsilon(a, b, c, d) \geq 0$ is achieved by restricting the range of ε to be $(-m, m)$ with m small enough, and (ii) summation $\int p_\varepsilon(a, b, c, d) da db dc dd = 1 + \varepsilon E[f] = 1$ for all ε , because any $f \in \mathcal{H}$ must have zero-mean, by definition of $\mathcal{H}$. This parametric model contain the truth p_0 at $\varepsilon = 0$. Moreover, the score function at p_0 is $\frac{\partial}{\partial \varepsilon} \log p_\varepsilon|_{\varepsilon=0} = f$. Finally, if we can show p_ε satisfies the DAG's constraints, then $\mathcal{P}_\varepsilon$ is a parametric submodel of the DAG's semiparametric model, and hence the chosen f must be in the tangent space $\mathcal{T}$.

We will directly check $p_\varepsilon(c \mid b, a) = p_\varepsilon(c \mid a)$ and $p_\varepsilon(d \mid c, b, a) = p_\varepsilon(d \mid c, b)$ to show that p_ε satisfies the constraints of the DAG model. This will conclude that any function in that subspace is a valid score function of this DAG model. Note that at the truth p_0 , the constraint $D \perp\!\!\!\perp A \mid C, B$ implies $p_0(d \mid c, b, a) = p_0(d \mid c, b)$, and the constraint $C \perp\!\!\!\perp B \mid A$ implies $p_0(c \mid b, a) = p_0(c \mid a)$ and $p_0(b \mid c, a) = p_0(b \mid a)$. We will use these equalities repeatedly in the proof. Finally, expectations are taken at the truth p_0 .

First case $\mathcal{T}_A \subseteq \mathcal{T}$: Pick any $f = \mathbb{E}[h \mid a] \in \mathcal{T}_A$. Then

$$p_\varepsilon(c, b, a) = \int p_0(d, c, b, a)(1 + \varepsilon \mathbb{E}[h \mid a]) dd = p_0(c, b, a)(1 + \varepsilon \mathbb{E}[h \mid a]) \quad (43)$$

Therefore, by definition of conditional distribution

$$\begin{aligned} p_\varepsilon(c \mid b, a) &= \frac{p_\varepsilon(c, b, a)}{p_\varepsilon(b, a)} = p_0(c \mid b, a) \\ p_\varepsilon(c \mid a) &= \frac{p_\varepsilon(c, a)}{p_\varepsilon(a)} = p_0(c \mid a) \\ p_\varepsilon(d \mid c, b, a) &= \frac{p_\varepsilon(d, c, b, a)}{p_\varepsilon(c, b, a)} = p_0(d \mid c, b, a) \end{aligned} \quad (44)$$

This shows $p_\varepsilon(c \mid b, a) = p_0(c \mid b, a) = p_0(c \mid a) = p_\varepsilon(c \mid a)$. Next

$$\begin{aligned} p_\varepsilon(d, c, b) &= \int p_0(d, c, b) p_0(a \mid c, b) (1 + \varepsilon \mathbb{E}[h \mid a]) da \quad (A \perp\!\!\!\perp D \mid C, B) \\ &= p_0(d, c, b) \left(1 + \varepsilon \int \mathbb{E}[h \mid a] p_0(a \mid c, b) da \right) \\ p_\varepsilon(c, b) &= p_0(c, b) \left(1 + \varepsilon \int \mathbb{E}[h \mid a] p_0(a \mid c, b) da \right) \\ \Rightarrow p_\varepsilon(d \mid c, b) &= \frac{p_\varepsilon(d, c, b)}{p_\varepsilon(c, b)} = p_0(d \mid c, b) \end{aligned} \quad (45)$$

This shows $p_\varepsilon(d \mid c, b, a) = p_0(d \mid c, b, a) = p_0(d \mid c, b) = p_\varepsilon(d \mid c, b)$. Hence $\mathcal{T}_A \subseteq \mathcal{T}$.

Second case $\mathcal{T}_{B|A} \subseteq \mathcal{T}$: Pick any $f = \mathbb{E}[h | b, a] - \mathbb{E}[h | a]$. Then

$$\begin{aligned} p_\varepsilon(c, b, a) &= p_0(c, b, a)(1 + \varepsilon\mathbb{E}[h | b, a] - \varepsilon\mathbb{E}[h | a]) \\ p_\varepsilon(b, a) &= p_0(b, a)(1 + \varepsilon\mathbb{E}[h | b, a] - \varepsilon\mathbb{E}[h | a]) \\ p_\varepsilon(c, a) &= p_0(c, a) \int p_0(b | c, a)(1 + \varepsilon\mathbb{E}[h | b, a] - \varepsilon\mathbb{E}[h | a])db \\ &= p_0(c, a) \int p_0(b | a)(1 + \varepsilon\mathbb{E}[h | b, a] - \varepsilon\mathbb{E}[h | a])db \quad (C \perp\!\!\!\perp B | A) \\ &= p_0(c, a) \end{aligned} \tag{46}$$

Therefore, we get $p_\varepsilon(c | b, a) = p_0(c | b, a) = p_0(c | a) = p_\varepsilon(c | a)$. Next,

$$\begin{aligned} p_\varepsilon(d, c, b) &= p_0(d, c, b) \int p_0(a | c, b)(1 + \varepsilon\mathbb{E}[h | b, a] - \varepsilon\mathbb{E}[h | a])da \quad (A \perp\!\!\!\perp D | C, B) \\ p_\varepsilon(c, b) &= \int p_0(c, b, a)(1 + \varepsilon\mathbb{E}[h | b, a] - \varepsilon\mathbb{E}[h | a])da \\ &= p_0(c, b) \int p_0(a | c, b)(1 + \varepsilon\mathbb{E}[h | b, a] - \varepsilon\mathbb{E}[h | a])da \end{aligned} \tag{47}$$

This shows $p_\varepsilon(d | c, b, a) = p_0(d | c, b, a) = p_0(d | c, b) = p_\varepsilon(d | c, b)$. Hence $\mathcal{T}_{B|A} \subseteq \mathcal{T}$.

Third case $\mathcal{T}_{C|A} \subseteq \mathcal{T}$: Similar to $\mathcal{T}_{B|A} \subseteq \mathcal{T}$.

Fourth case $\mathcal{T}_{D|CB} \subseteq \mathcal{T}$: Pick $f = \mathbb{E}[h | d, c, b] - \mathbb{E}[h | c, b]$. Then

$$\begin{aligned} p_\varepsilon(c, b, a) &= p_0(c, b, a) \int p_0(d | c, b)(1 + \varepsilon\mathbb{E}[h | d, c, b] - \varepsilon\mathbb{E}[h | c, b])dd \quad (D \perp\!\!\!\perp A | C, B) \\ &= p_0(c, b, a) \end{aligned} \tag{48}$$

This show $p_\varepsilon(c | b, a) = p_0(c | b, a) = p_0(c | a) = p_\varepsilon(c | a)$. Moreover, $p_\varepsilon(d | c, b, a) = p_0(d | c, b, a) \times \int f p_0(d | c, b)dd$. Next

$$\begin{aligned} p_\varepsilon(d, c, b) &= p_0(d, c, b) \int p_0(a | d, c, b)(1 + \varepsilon\mathbb{E}[h | d, c, b] - \varepsilon\mathbb{E}[h | c, b])da \\ &= p_0(d, c, b)(1 + \varepsilon\mathbb{E}[h | d, c, b] - \varepsilon\mathbb{E}[h | c, b]) \\ p_\varepsilon(c, b) &= p_0(c, b) \int p_0(d | c, b)(1 + \varepsilon\mathbb{E}[h | d, c, b] - \varepsilon\mathbb{E}[h | c, b])dd \\ &= p_0(c, b) \end{aligned} \tag{49}$$

This shows $p_\varepsilon(d | c, b) = p_0(d | c, b) \times \int f p_0(d | c, b)dd$. Therefore $p_\varepsilon(d | c, b, a) = p_\varepsilon(d | c, b)$. Hence $\mathcal{T}_{D|CB} \subseteq \mathcal{T}$.

The proof is established because $\mathcal{T}_A, \mathcal{T}_{B|A}, \mathcal{T}_{C|A}, \mathcal{T}_{D|CB} \subseteq \mathcal{T}$ implies $\mathcal{T}_A \oplus \mathcal{T}_{B|A} \oplus \mathcal{T}_{C|A} \oplus \mathcal{T}_{D|CB} \subseteq \mathcal{T}$, and the equality follows. $\square$

We reproduce a special case of the well-known result about DAG model's tangent space and its orthogonal complement [Tsiatis, 2006, Bhattacharya et al., 2022, Rotnitzky and Smucler, 2020] in the following lemma.

Lemma 1. Suppose $\mathcal{P}$ is a single independence DAG model over variables $\mathbf{V}$, defined by a single constraint $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$. The model's tangent space $\mathcal{T}$ and its orthogonal complement $\mathcal{T}^\perp$ are, respectively

$$\begin{aligned} \mathcal{T} &= \{h - \mathbb{E}[h | \mathbf{x}, \mathbf{y}, \mathbf{z}] + \mathbb{E}[h | \mathbf{x}, \mathbf{z}] - \mathbb{E}[h | \mathbf{z}] \\ &\quad + \mathbb{E}[h | \mathbf{y}, \mathbf{z}] : \forall h \in \mathcal{H}\}, \\ \mathcal{T}^\perp &= \{\mathbb{E}[h | \mathbf{x}, \mathbf{y}, \mathbf{z}] - \mathbb{E}[h | \mathbf{y}, \mathbf{z}] \\ &\quad - \mathbb{E}[h | \mathbf{x}, \mathbf{z}] + \mathbb{E}[h | \mathbf{z}] : \forall h \in \mathcal{H}\}. \end{aligned} \tag{11}$$

Proof. Let $\mathbf{W} = \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z})$. For any parametric submodel $\{p_\varepsilon(\mathbf{v})\}$ of $\mathcal{P}$, its elements factorize as

$$p_\varepsilon(\mathbf{v}) = p_\varepsilon(\mathbf{w} | \mathbf{x}, \mathbf{y}, \mathbf{z})p_\varepsilon(\mathbf{x} | \mathbf{z})p_\varepsilon(\mathbf{y} | \mathbf{z})p_\varepsilon(\mathbf{z}) \tag{50}$$

The total score function is $p_\varepsilon(\mathbf{v}) = \frac{\partial}{\partial \varepsilon} p_\varepsilon(\mathbf{v})|_{\varepsilon=0}$. By Lemma 5,

$$s(\mathbf{w} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) := \frac{\partial}{\partial \varepsilon} p_\varepsilon(\mathbf{w} \mid \mathbf{x}, \mathbf{y}, \mathbf{z})|_{\varepsilon=0} = s(\mathbf{V}) - \mathbb{E}[s(\mathbf{V}) \mid \mathbf{x}, \mathbf{y}, \mathbf{z}] \quad (51)$$

$$s(\mathbf{x} \mid \mathbf{z}) := \frac{\partial}{\partial \varepsilon} p_\varepsilon(\mathbf{x} \mid \mathbf{z})|_{\varepsilon=0} = \mathbb{E}[s(\mathbf{V}) \mid \mathbf{x}, \mathbf{z}] - \mathbb{E}[s(\mathbf{V}) \mid \mathbf{z}] \quad (52)$$

$$s(\mathbf{y} \mid \mathbf{z}) := \frac{\partial}{\partial \varepsilon} p_\varepsilon(\mathbf{y} \mid \mathbf{z})|_{\varepsilon=0} = \mathbb{E}[s(\mathbf{V}) \mid \mathbf{y}, \mathbf{z}] - \mathbb{E}[s(\mathbf{V}) \mid \mathbf{z}] \quad (53)$$

$$s(\mathbf{z}) := \frac{\partial}{\partial \varepsilon} p_\varepsilon(\mathbf{z})|_{\varepsilon=0} = \mathbb{E}[s(\mathbf{V}) \mid \mathbf{z}] \quad (54)$$

Define the following subspaces

$$\mathcal{T}_{\mathbf{W} \mid \mathbf{X} \mathbf{Y} \mathbf{Z}} := \{h - \mathbb{E}[h \mid \mathbf{x}, \mathbf{y}, \mathbf{z}] : h \in \mathcal{H}\} \quad (55)$$

$$\mathcal{T}_{\mathbf{X} \mid \mathbf{Z}} := \{\mathbb{E}[h \mid \mathbf{x}, \mathbf{z}] - \mathbb{E}[h \mid \mathbf{z}] : h \in \mathcal{H}\} \quad (56)$$

$$\mathcal{T}_{\mathbf{Y} \mid \mathbf{Z}} := \{\mathbb{E}[h \mid \mathbf{y}, \mathbf{z}] - \mathbb{E}[h \mid \mathbf{z}] : h \in \mathcal{H}\} \quad (57)$$

$$\mathcal{T}_{\mathbf{Z}} := \{\mathbb{E}[h \mid \mathbf{z}] : h \in \mathcal{H}\} \quad (58)$$

It is evident that $s(\mathbf{w} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{T}_{\mathbf{W} \mid \mathbf{X} \mathbf{Y} \mathbf{Z}}$, $s(\mathbf{x} \mid \mathbf{z}) \in \mathcal{T}_{\mathbf{X} \mid \mathbf{Z}}$, $s(\mathbf{y} \mid \mathbf{z}) \in \mathcal{T}_{\mathbf{Y} \mid \mathbf{Z}}$ and $s(\mathbf{z}) \in \mathcal{T}_{\mathbf{Z}}$. By Lemma 6, these subspaces are orthogonal. Moreover, due to Equation 50

$$s(\mathbf{v}) = s(\mathbf{w} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) + s(\mathbf{x} \mid \mathbf{z}) + s(\mathbf{y} \mid \mathbf{z}) + s(\mathbf{z}) \quad (59)$$

Then $s(\mathbf{v}) \in \mathcal{T}_{\mathbf{W} \mid \mathbf{X} \mathbf{Y} \mathbf{Z}} \oplus \mathcal{T}_{\mathbf{X} \mid \mathbf{Z}} \oplus \mathcal{T}_{\mathbf{Y} \mid \mathbf{Z}} \oplus \mathcal{T}_{\mathbf{Z}}$. This means the tangent space of this model is $\mathcal{T} \subseteq \mathcal{T}_{\mathbf{W} \mid \mathbf{X} \mathbf{Y} \mathbf{Z}} \oplus \mathcal{T}_{\mathbf{X} \mid \mathbf{Z}} \oplus \mathcal{T}_{\mathbf{Y} \mid \mathbf{Z}} \oplus \mathcal{T}_{\mathbf{Z}}$.

Next, we will show that each of the subspace is in $\mathcal{T}$. To do this, we must show that any function in these subspaces can be used to construct parametric submodel of $\mathcal{P}$.

Pick any $f = h - \mathbb{E}[h \mid \mathbf{x}, \mathbf{y}, \mathbf{z}] \in \mathcal{T}_{\mathbf{W} \mid \mathbf{X} \mathbf{Y} \mathbf{Z}}$ and construct a parametric model by $p_\varepsilon(\mathbf{v}) = p_0(\mathbf{v})(1 + \varepsilon f(\mathbf{v}))$. Since $\mathbb{E}[f] = 0$, p_ε sums to 1. Moreover, the parameter $\varepsilon \in (-\delta, \delta)$ and δ is small enough so that $p_\varepsilon(\mathbf{v}) \geq 0$ at all $\mathbf{v}$. Then p_ε is a valid probability distribution. This parametric model contains p_0 at $\varepsilon = 0$, and the score function at p_0 is $\frac{\partial}{\partial \varepsilon} \log p_\varepsilon(\mathbf{v})|_{\varepsilon=0} = f$. What left to show is that the constraint holds for all p_ε . Indeed, for any p_ε ,

$$p_\varepsilon(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p_0(\mathbf{x}, \mathbf{y}, \mathbf{z}) \left(1 + \varepsilon \int f(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}) p_0(\mathbf{w} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) d\mathbf{w} \right) = p_0(\mathbf{x}, \mathbf{y}, \mathbf{z}), \quad (60)$$

since $\mathbb{E}[f \mid \mathbf{x}, \mathbf{y}, \mathbf{z}] = 0$. Therefore the constraint $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ holds for p_ε , $p_\varepsilon(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) = p_\varepsilon(\mathbf{x} \mid \mathbf{z})$, because it holds for p_0 . This means $\{p_\varepsilon(\mathbf{v}) : \varepsilon \in (-\delta, \delta)\}$ is also a parametric submodel of $\mathcal{P}$, hence $f \in \mathcal{T}$ and therefore $\mathcal{T}_{\mathbf{W} \mid \mathbf{X} \mathbf{Y} \mathbf{Z}} \subseteq \mathcal{T}$.

Pick any $f = \mathbb{E}[h \mid \mathbf{x}, \mathbf{z}] - \mathbb{E}[h \mid \mathbf{z}] \in \mathcal{T}_{\mathbf{X} \mid \mathbf{Z}}$ and construct a parametric model by $p_\varepsilon(\mathbf{v}) = p_0(\mathbf{v})(1 + \varepsilon f(\mathbf{v}))$. By the same reasoning, all we need to show is the constraint holds for all $p_\varepsilon(\mathbf{v})$.

$$\begin{aligned} p_\varepsilon(\mathbf{x}, \mathbf{y}, \mathbf{z}) &= p_0(\mathbf{x}, \mathbf{y}, \mathbf{z})(1 + \varepsilon \mathbb{E}[f \mid \mathbf{x}, \mathbf{y}, \mathbf{z}]) = p_0(\mathbf{x}, \mathbf{y}, \mathbf{z})(1 + \varepsilon \mathbb{E}[h \mid \mathbf{x}, \mathbf{z}] - \varepsilon \mathbb{E}[h \mid \mathbf{z}]) \\ p_\varepsilon(\mathbf{y}, \mathbf{z}) &= p_0(\mathbf{y}, \mathbf{z})(1 + \varepsilon \mathbb{E}[f \mid \mathbf{y}, \mathbf{z}]) = p_0(\mathbf{y}, \mathbf{z}) \left(1 + \varepsilon \int \mathbb{E}[h \mid \mathbf{x}, \mathbf{z}] p_0(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) d\mathbf{x} - \varepsilon \mathbb{E}[h \mid \mathbf{z}] \right) \\ &= p_0(\mathbf{y}, \mathbf{z}) \left(1 + \varepsilon \int \mathbb{E}[h \mid \mathbf{x}, \mathbf{z}] p_0(\mathbf{x} \mid \mathbf{z}) d\mathbf{x} - \varepsilon \mathbb{E}[h \mid \mathbf{z}] \right) \quad (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \text{ at } p_0) \\ &= p_0(\mathbf{y}, \mathbf{z}) \\ p_\varepsilon(\mathbf{x}, \mathbf{z}) &= p_0(\mathbf{x}, \mathbf{z})(1 + \varepsilon \mathbb{E}[f \mid \mathbf{x}, \mathbf{z}]) = p_0(\mathbf{x}, \mathbf{z})(1 + \varepsilon \mathbb{E}[h \mid \mathbf{x}, \mathbf{z}] - \varepsilon \mathbb{E}[h \mid \mathbf{z}]) \\ p_\varepsilon(\mathbf{z}) &= p_0(\mathbf{z})(1 + \varepsilon \mathbb{E}[f \mid \mathbf{z}]) = p_0(\mathbf{z}) \quad (\text{since } \mathbb{E}[f \mid \mathbf{z}] = 0) \end{aligned} \quad (61)$$

Therefore,

$$\begin{aligned} p_\varepsilon(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) &:= \frac{p_\varepsilon(\mathbf{x}, \mathbf{y}, \mathbf{z})}{p_\varepsilon(\mathbf{y}, \mathbf{z})} = p_0(\mathbf{x} \mid \mathbf{y}, \mathbf{z})(1 + \varepsilon \mathbb{E}[h \mid \mathbf{x}, \mathbf{z}] - \varepsilon \mathbb{E}[h \mid \mathbf{z}]) \\ &= p_0(\mathbf{x} \mid \mathbf{z})(1 + \varepsilon \mathbb{E}[h \mid \mathbf{x}, \mathbf{z}] - \varepsilon \mathbb{E}[h \mid \mathbf{z}]) = \frac{p_\varepsilon(\mathbf{x}, \mathbf{z})}{p_\varepsilon(\mathbf{z})} = p_\varepsilon(\mathbf{x} \mid \mathbf{z}) \end{aligned} \quad (62)$$

Hence, the constraint holds for all p_ε , so $\{p_\varepsilon\}$ is a parametric submodel of $\mathcal{P}$, and therefore $f \in \mathcal{T}_{\mathbf{X} \mid \mathbf{Z}} \subseteq \mathcal{T}$.

Pick any $f = \mathbb{E}[h \mid \mathbf{y}, \mathbf{z}] - \mathbb{E}[h \mid \mathbf{z}] \in \mathcal{T}_{\mathbf{X}|\mathbf{Z}}$ and construct a parametric model by $p_\varepsilon(\mathbf{v}) = p_0(\mathbf{v})(1 + \varepsilon f(\mathbf{v}))$. By the same reasoning, all we need to show is the constraint holds for all $p_\varepsilon(\mathbf{v})$.

$$\begin{aligned}
p_\varepsilon(\mathbf{x}, \mathbf{y}, \mathbf{z}) &= p_0(\mathbf{x}, \mathbf{y}, \mathbf{z})(1 + \varepsilon \mathbb{E}[f \mid \mathbf{x}, \mathbf{y}, \mathbf{z}]) = p_0(\mathbf{x}, \mathbf{y}, \mathbf{z})(1 + \varepsilon \mathbb{E}[h \mid \mathbf{y}, \mathbf{z}] - \varepsilon \mathbb{E}[h \mid \mathbf{z}]) \\
p_\varepsilon(\mathbf{y}, \mathbf{z}) &= p_0(\mathbf{y}, \mathbf{z})(1 + \varepsilon \mathbb{E}[f \mid \mathbf{y}, \mathbf{z}]) = p_0(\mathbf{y}, \mathbf{z})(1 + \varepsilon \mathbb{E}[h \mid \mathbf{y}, \mathbf{z}] - \varepsilon \mathbb{E}[h \mid \mathbf{z}]) \\
p_\varepsilon(\mathbf{x}, \mathbf{z}) &= p_0(\mathbf{x}, \mathbf{z})(1 + \varepsilon \mathbb{E}[f \mid \mathbf{x}, \mathbf{z}]) = p_0(\mathbf{x}, \mathbf{z}) \left(1 + \varepsilon \int \mathbb{E}[h \mid \mathbf{y}, \mathbf{z}] p_0(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) d\mathbf{y} - \varepsilon \mathbb{E}[h \mid \mathbf{z}] \right) \\
&= p_0(\mathbf{x}, \mathbf{z}) \left(1 + \varepsilon \int \mathbb{E}[h \mid \mathbf{y}, \mathbf{z}] p_0(\mathbf{y} \mid \mathbf{z}) d\mathbf{x} - \varepsilon \mathbb{E}[h \mid \mathbf{z}] \right) \quad (\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z} \text{ at } p_0) \\
&= p_0(\mathbf{x}, \mathbf{z}) \\
p_\varepsilon(\mathbf{z}) &= p_0(\mathbf{z})(1 + \varepsilon \mathbb{E}[f \mid \mathbf{z}]) = p_0(\mathbf{z}) \quad (\text{since } \mathbb{E}[f \mid \mathbf{z}] = 0)
\end{aligned} \tag{63}$$

Therefore,

$$p_\varepsilon(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) := \frac{p_\varepsilon(\mathbf{x}, \mathbf{y}, \mathbf{z})}{p_\varepsilon(\mathbf{y}, \mathbf{z})} = p_0(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) = p_0(\mathbf{x} \mid \mathbf{z}) = p_\varepsilon(\mathbf{x} \mid \mathbf{z}) \tag{64}$$

Hence, the constraint holds for all p_ε , so $\{p_\varepsilon\}$ is a parametric submodel of $\mathcal{P}$, and therefore $f \in \mathcal{T}_{\mathbf{Y}|\mathbf{Z}} \subseteq \mathcal{T}$.

Pick any $f = \mathbb{E}[h \mid \mathbf{z}] \in \mathcal{T}_{\mathbf{Z}}$ and construct a parametric model by $p_\varepsilon(\mathbf{v}) = p_0(\mathbf{v})(1 + \varepsilon f(\mathbf{v}))$. Then

$$\begin{aligned}
p_\varepsilon(\mathbf{x}, \mathbf{y}, \mathbf{z}) &= p_0(\mathbf{x}, \mathbf{y}, \mathbf{z})(1 + \varepsilon \mathbb{E}[f \mid \mathbf{x}, \mathbf{y}, \mathbf{z}]) = p_0(\mathbf{x}, \mathbf{y}, \mathbf{z})(1 + \varepsilon \mathbb{E}[h \mid \mathbf{z}]) \\
p_\varepsilon(\mathbf{y}, \mathbf{z}) &= p_0(\mathbf{y}, \mathbf{z})(1 + \varepsilon \mathbb{E}[f \mid \mathbf{y}, \mathbf{z}]) = p_0(\mathbf{y}, \mathbf{z})(1 + \varepsilon \mathbb{E}[h \mid \mathbf{z}]) \\
p_\varepsilon(\mathbf{x}, \mathbf{z}) &= p_0(\mathbf{x}, \mathbf{z})(1 + \varepsilon \mathbb{E}[f \mid \mathbf{x}, \mathbf{z}]) = p_0(\mathbf{x}, \mathbf{z})(1 + \varepsilon \mathbb{E}[h \mid \mathbf{z}]) \\
p_\varepsilon(\mathbf{z}) &= p_0(\mathbf{z})(1 + \varepsilon \mathbb{E}[f \mid \mathbf{z}]) = p_0(\mathbf{z})(1 + \varepsilon \mathbb{E}[h \mid \mathbf{z}])
\end{aligned} \tag{65}$$

Therefore,

$$p_\varepsilon(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) := \frac{p_\varepsilon(\mathbf{x}, \mathbf{y}, \mathbf{z})}{p_\varepsilon(\mathbf{y}, \mathbf{z})} = p_0(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) = p_0(\mathbf{x} \mid \mathbf{z}) = p_\varepsilon(\mathbf{x} \mid \mathbf{z}) \tag{66}$$

Hence, the constraint holds for all p_ε , so $\{p_\varepsilon\}$ is a parametric submodel of $\mathcal{P}$, and therefore $f \in \mathcal{T}_{\mathbf{Z}} \subseteq \mathcal{T}$.

These results imply $\mathcal{T}_{\mathbf{W}|\mathbf{X}\mathbf{Y}\mathbf{Z}} \oplus \mathcal{T}_{\mathbf{X}|\mathbf{Z}} \oplus \mathcal{T}_{\mathbf{Y}|\mathbf{Z}} \oplus \mathcal{T}_{\mathbf{Z}} \subseteq \mathcal{T}$, and with previous result, we get $\mathcal{T} = \mathcal{T}_{\mathbf{W}|\mathbf{X}\mathbf{Y}\mathbf{Z}} \oplus \mathcal{T}_{\mathbf{X}|\mathbf{Z}} \oplus \mathcal{T}_{\mathbf{Y}|\mathbf{Z}} \oplus \mathcal{T}_{\mathbf{Z}}$.

Define the following operators

$$\Pi(h \mid \mathcal{T}_{\mathbf{W}|\mathbf{X}\mathbf{Y}\mathbf{Z}}) = h - \mathbb{E}[h \mid \mathbf{x}, \mathbf{y}, \mathbf{z}] \tag{67}$$

$$\Pi(h \mid \mathcal{T}_{\mathbf{X}|\mathbf{Z}}) = \mathbb{E}[h \mid \mathbf{x}, \mathbf{z}] - \mathbb{E}[h \mid \mathbf{z}] \tag{68}$$

$$\Pi(h \mid \mathcal{T}_{\mathbf{Y}|\mathbf{Z}}) = \mathbb{E}[h \mid \mathbf{y}, \mathbf{z}] - \mathbb{E}[h \mid \mathbf{z}] \tag{69}$$

$$\Pi(h \mid \mathcal{T}_{\mathbf{Z}}) = \mathbb{E}[h \mid \mathbf{z}] \tag{70}$$

It's an exercise left to the reader to show that $\Pi(\cdot \mid \mathcal{T}_{\mathbf{W}|\mathbf{X}\mathbf{Y}\mathbf{Z}}), \Pi(\cdot \mid \mathcal{T}_{\mathbf{X}|\mathbf{Z}}), \Pi(\cdot \mid \mathcal{T}_{\mathbf{Y}|\mathbf{Z}}), \Pi(\cdot \mid \mathcal{T}_{\mathbf{Z}})$ are the orthogonal projection operators onto $\mathcal{T}_{\mathbf{W}|\mathbf{X}\mathbf{Y}\mathbf{Z}}, \mathcal{T}_{\mathbf{X}|\mathbf{Z}}, \mathcal{T}_{\mathbf{Y}|\mathbf{Z}}, \mathcal{T}_{\mathbf{Z}}$, respectively. In particular, to show that $\Pi(\cdot)$ is a projection operator onto a subspace $\mathcal{S}$, one only needs to show: (i) $\mathcal{S} = \{\Pi(h) : h \in \mathcal{H}\}$ and (ii) $f - \Pi(f)$ and $\Pi(g)$ are orthogonal, for all $f, g \in \mathcal{H}$.

Define the following operator

$$\begin{aligned}
\Pi(h) &= \Pi(h \mid \mathcal{T}_{\mathbf{W}|\mathbf{X}\mathbf{Y}\mathbf{Z}}) + \Pi(h \mid \mathcal{T}_{\mathbf{X}|\mathbf{Z}}) + \Pi(h \mid \mathcal{T}_{\mathbf{Y}|\mathbf{Z}}) + \Pi(h \mid \mathcal{T}_{\mathbf{Z}}) \\
&= h - \mathbb{E}[h \mid \mathbf{x}, \mathbf{y}, \mathbf{z}] + \mathbb{E}[h \mid \mathbf{x}, \mathbf{z}] - \mathbb{E}[h \mid \mathbf{z}] + \mathbb{E}[h \mid \mathbf{y}, \mathbf{z}]
\end{aligned} \tag{71}$$

We will show that Π is the projection operator onto the tangent space $\mathcal{T}$.

First, we will show $\mathcal{T} = \{\Pi(h) : \forall h \in \mathcal{H}\}$. For the first direction, $\Pi(h) \in \mathcal{T}$ for all $h \in \mathcal{H}$, obviously. For the other direction, since $\mathcal{T} = \mathcal{T}_{\mathbf{W}|\mathbf{X}\mathbf{Y}\mathbf{Z}} \oplus \mathcal{T}_{\mathbf{X}|\mathbf{Z}} \oplus \mathcal{T}_{\mathbf{Y}|\mathbf{Z}} \oplus \mathcal{T}_{\mathbf{Z}}$, any $f \in \mathcal{T}$ can be written as, for some $h_1, h_2, h_3, h_4 \in \mathcal{H}$,

$$f = h_1 - \mathbb{E}[h_1 \mid \mathbf{x}, \mathbf{y}, \mathbf{z}] + \mathbb{E}[h_2 \mid \mathbf{x}, \mathbf{z}] - \mathbb{E}[h_2 \mid \mathbf{z}] + \mathbb{E}[h_3 \mid \mathbf{y}, \mathbf{z}] - \mathbb{E}[h_3 \mid \mathbf{z}] + \mathbb{E}[h_4 \mid \mathbf{z}]. \tag{72}$$

Because the subspaces $\mathcal{T}_{\mathbf{W}|\mathbf{X}\mathbf{Y}\mathbf{Z}}, \mathcal{T}_{\mathbf{X}|\mathbf{Z}}, \mathcal{T}_{\mathbf{Y}|\mathbf{Z}}, \mathcal{T}_{\mathbf{Z}}$ are orthogonal,

$$\begin{aligned}\Pi(f | \mathcal{T}_{\mathbf{W}|\mathbf{X}\mathbf{Y}\mathbf{Z}}) &= h_1 - \mathbb{E}[h_1 | \mathbf{x}, \mathbf{y}, \mathbf{z}] \\ \Pi(f | \mathcal{T}_{\mathbf{X}|\mathbf{Z}}) &= \mathbb{E}[h_2 | \mathbf{x}, \mathbf{z}] - \mathbb{E}[h_2 | \mathbf{z}] \\ \Pi(f | \mathcal{T}_{\mathbf{Y}|\mathbf{Z}}) &= \mathbb{E}[h_3 | \mathbf{y}, \mathbf{z}] - \mathbb{E}[h_3 | \mathbf{z}] \\ \Pi(f | \mathcal{T}_{\mathbf{Z}}) &= \mathbb{E}[h_4 | \mathbf{z}].\end{aligned}\tag{73}$$

Therefore, $f = \Pi(f)$, hence $\mathcal{T} \subseteq \{\Pi(h) : h \in \mathcal{H}\}$. This implies $\mathcal{T} = \{\Pi(h) : h \in \mathcal{H}\}$.

Second, we will show that $f - \Pi(f)$ and $\Pi(g)$ are orthogonal, for all $f, g \in \mathcal{H}$.

$$\begin{aligned}\mathbb{E} \left[f - \Pi(f), \Pi(g) \right] &= \mathbb{E} \left[f \Pi(g) \right] - \mathbb{E} \left[\Pi(f) \Pi(g) \right] \\ &= \mathbb{E} \left[\left(\mathbb{E}[f | \mathbf{X}, \mathbf{Y}, \mathbf{Z}] - \mathbb{E}[f | \mathbf{X}, \mathbf{Z}] + \mathbb{E}[f | \mathbf{Z}] - \mathbb{E}[f | \mathbf{Y}, \mathbf{Z}] \right) \times \left(g - \mathbb{E}[g | \mathbf{X}, \mathbf{Y}, \mathbf{Z}] + \mathbb{E}[g | \mathbf{X}, \mathbf{Z}] - \mathbb{E}[g | \mathbf{Z}] + \mathbb{E}[g | \mathbf{Y}, \mathbf{Z}] \right) \right] \\ &= \mathbb{E} \left[\left(\mathbb{E}[f | \mathbf{X}, \mathbf{Y}, \mathbf{Z}] - \mathbb{E}[f | \mathbf{X}, \mathbf{Z}] + \mathbb{E}[f | \mathbf{Z}] - \mathbb{E}[f | \mathbf{Y}, \mathbf{Z}] \right) \times \left(\mathbb{E}[g | \mathbf{X}, \mathbf{Z}] - \mathbb{E}[g | \mathbf{Z}] + \mathbb{E}[g | \mathbf{Y}, \mathbf{Z}] \right) \right] \\ &= \mathbb{E} \left[\mathbb{E}[f | \mathbf{X}, \mathbf{Y}, \mathbf{Z}] \mathbb{E}[g | \mathbf{X}, \mathbf{Z}] \right] - \mathbb{E} \left[\mathbb{E}[f | \mathbf{X}, \mathbf{Z}] \mathbb{E}[g | \mathbf{X}, \mathbf{Z}] \right] + \mathbb{E} \left[\mathbb{E}[f | \mathbf{Z}] \mathbb{E}[g | \mathbf{X}, \mathbf{Z}] \right] - \mathbb{E} \left[\mathbb{E}[f | \mathbf{Y}, \mathbf{Z}] \mathbb{E}[g | \mathbf{X}, \mathbf{Z}] \right] \\ &\quad - \mathbb{E} \left[\mathbb{E}[f | \mathbf{X}, \mathbf{Y}, \mathbf{Z}] \mathbb{E}[g | \mathbf{Z}] \right] + \mathbb{E} \left[\mathbb{E}[f | \mathbf{X}, \mathbf{Z}] \mathbb{E}[g | \mathbf{Z}] \right] - \mathbb{E} \left[\mathbb{E}[f | \mathbf{Z}] \mathbb{E}[g | \mathbf{Z}] \right] + \mathbb{E} \left[\mathbb{E}[f | \mathbf{Y}, \mathbf{Z}] \mathbb{E}[g | \mathbf{Z}] \right] \\ &\quad + \mathbb{E} \left[\mathbb{E}[f | \mathbf{X}, \mathbf{Y}, \mathbf{Z}] \mathbb{E}[g | \mathbf{Y}, \mathbf{Z}] \right] - \mathbb{E} \left[\mathbb{E}[f | \mathbf{X}, \mathbf{Z}] \mathbb{E}[g | \mathbf{Y}, \mathbf{Z}] \right] + \mathbb{E} \left[\mathbb{E}[f | \mathbf{Z}] \mathbb{E}[g | \mathbf{Y}, \mathbf{Z}] \right] - \mathbb{E} \left[\mathbb{E}[f | \mathbf{Y}, \mathbf{Z}] \mathbb{E}[g | \mathbf{Y}, \mathbf{Z}] \right] \\ &= \mathbb{E} \left[\mathbb{E}[f | \mathbf{X}, \mathbf{Z}] \mathbb{E}[g | \mathbf{X}, \mathbf{Z}] \right] - \mathbb{E} \left[\mathbb{E}[f | \mathbf{X}, \mathbf{Z}] \mathbb{E}[g | \mathbf{X}, \mathbf{Z}] \right] + \mathbb{E} \left[\mathbb{E}[f | \mathbf{Z}] \mathbb{E}[g | \mathbf{Z}] \right] - \mathbb{E} \left[\mathbb{E}[f | \mathbf{Y}, \mathbf{Z}] \mathbb{E}[g | \mathbf{X}, \mathbf{Z}] \right] \\ &\quad - \mathbb{E} \left[\mathbb{E}[f | \mathbf{Z}] \mathbb{E}[g | \mathbf{Z}] \right] + \mathbb{E} \left[\mathbb{E}[f | \mathbf{Z}] \mathbb{E}[g | \mathbf{Z}] \right] - \mathbb{E} \left[\mathbb{E}[f | \mathbf{Z}] \mathbb{E}[g | \mathbf{Z}] \right] + \mathbb{E} \left[\mathbb{E}[f | \mathbf{Z}] \mathbb{E}[g | \mathbf{Z}] \right] \\ &\quad + \mathbb{E} \left[\mathbb{E}[f | \mathbf{Y}, \mathbf{Z}] \mathbb{E}[g | \mathbf{Y}, \mathbf{Z}] \right] - \mathbb{E} \left[\mathbb{E}[f | \mathbf{X}, \mathbf{Z}] \mathbb{E}[g | \mathbf{Y}, \mathbf{Z}] \right] + \mathbb{E} \left[\mathbb{E}[f | \mathbf{Z}] \mathbb{E}[g | \mathbf{Z}] \right] - \mathbb{E} \left[\mathbb{E}[f | \mathbf{Y}, \mathbf{Z}] \mathbb{E}[g | \mathbf{Y}, \mathbf{Z}] \right] \\ &= -\mathbb{E} \left[\mathbb{E}[f | \mathbf{Y}, \mathbf{Z}] \mathbb{E}[g | \mathbf{X}, \mathbf{Z}] \right] - \mathbb{E} \left[\mathbb{E}[f | \mathbf{X}, \mathbf{Z}] \mathbb{E}[g | \mathbf{Y}, \mathbf{Z}] \right] + 2 \mathbb{E} \left[\mathbb{E}[f | \mathbf{Z}] \mathbb{E}[g | \mathbf{Z}] \right]\end{aligned}\tag{74}$$

Due to the constraint

$$\begin{aligned}\mathbb{E} \left[\mathbb{E}[f | \mathbf{Y}, \mathbf{Z}] \mathbb{E}[g | \mathbf{X}, \mathbf{Z}] \right] &= \int \mathbb{E}[f | \mathbf{y}, \mathbf{z}] \mathbb{E}[g | \mathbf{x}, \mathbf{z}] p_0(\mathbf{x}, \mathbf{y}, \mathbf{z}) d\mathbf{x} d\mathbf{y} d\mathbf{z} \\ &= \int \mathbb{E}[f | \mathbf{y}, \mathbf{z}] \mathbb{E}[g | \mathbf{x}, \mathbf{z}] p_0(\mathbf{x} | \mathbf{z}) p_0(\mathbf{y} | \mathbf{z}) p_0(\mathbf{z}) d\mathbf{x} d\mathbf{y} d\mathbf{z} \\ &= \int \mathbb{E}[f | \mathbf{z}] \mathbb{E}[g | \mathbf{z}] p_0(\mathbf{z}) d\mathbf{z} \\ &= \mathbb{E} \left[\mathbb{E}[f | \mathbf{Z}] \mathbb{E}[g | \mathbf{Z}] \right]\end{aligned}\tag{75}$$

Similarly, $\mathbb{E} \left[\mathbb{E}[f | \mathbf{X}, \mathbf{Z}] \mathbb{E}[g | \mathbf{Y}, \mathbf{Z}] \right] = \mathbb{E} \left[\mathbb{E}[f | \mathbf{Z}] \mathbb{E}[g | \mathbf{Z}] \right]$. Therefore, the inner product vanishes,

$$\mathbb{E} \left[f - \Pi(f), \Pi(g) \right] = 0.$$

In conclusion, the operator defined by

$$\Pi(h | \mathcal{T}) = h - \mathbb{E}[h | \mathbf{x}, \mathbf{y}, \mathbf{z}] + \mathbb{E}[h | \mathbf{x}, \mathbf{z}] - \mathbb{E}[h | \mathbf{z}] + \mathbb{E}[h | \mathbf{y}, \mathbf{z}]\tag{76}$$

is the projection operator onto the tangent space of this model, hence $\mathcal{T} = \{\Pi(h | \mathcal{T}) : \forall h \in \mathcal{H}\}$. The operator onto its orthogonal complement is

$$\begin{aligned}\Pi(h | \mathcal{T}^\perp) &= h - \Pi(h | \mathcal{T}) \\ &= \mathbb{E}[h | \mathbf{x}, \mathbf{y}, \mathbf{z}] - \mathbb{E}[h | \mathbf{x}, \mathbf{z}] + \mathbb{E}[h | \mathbf{z}] - \mathbb{E}[h | \mathbf{y}, \mathbf{z}]\end{aligned}\tag{77}$$

and $\mathcal{T}^\perp = \{\Pi(h | \mathcal{T}^\perp) : \forall h \in \mathcal{H}\}$.

□

C PROOFS IN SECTION 4

Theorem 3. Suppose $\mathcal{P}$ is a general semi-parametric Markov model defined by $K \geq 1$ conditional independence constraints $\mathbf{X}_i \perp\!\!\!\perp \mathbf{Y}_i \mid \mathbf{Z}_i$, where $i = 1, \dots, K$. For each i , let $\mathcal{P}_i$ be the single independence DAG model defined by the i -th constraint. The orthogonal projection operator $\Pi(\cdot \mid \mathcal{T}_i^\perp) : \mathcal{H} \mapsto \mathcal{H}$ sends each function $h \in \mathcal{H}$ to the orthogonal complement $\mathcal{T}_i^\perp$ of the tangent space $\mathcal{T}_i$ of $\mathcal{P}_i$,

$$\Pi(h \mid \mathcal{T}_i^\perp)(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i) = \mathbb{E}[h|\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i] - \mathbb{E}[h|\mathbf{x}_i, \mathbf{z}_i] - \mathbb{E}[h|\mathbf{y}_i, \mathbf{z}_i] + \mathbb{E}[h|\mathbf{z}_i]. \quad (78)$$

The orthogonal complement of the tangent space of $\mathcal{P}$ is the direct sum of these orthogonal complements,

$$\mathcal{T}^\perp = \left\{ \sum_{i=1}^K \Pi(h_i \mid \mathcal{T}_i^\perp) : \forall h_1, \dots, h_K \in \mathcal{H} \right\}. \quad (79)$$

Proof. First, we will show $\mathcal{T} = \mathcal{T}_1 \cap \dots \cap \mathcal{T}_K$.

Let $\{p_\varepsilon(\mathbf{v}) : \varepsilon \in (-\delta, \delta)\}$ be any parametric submodel of $\mathcal{P}$, with total score function $s(\mathbf{v})$ at p_0 . For each i , $p_\varepsilon(\mathbf{v})$ satisfies the constraint $\mathbf{X}_i \perp\!\!\!\perp \mathbf{Y}_i \mid \mathbf{Z}_i$. Therefore, $\{p_\varepsilon(\mathbf{v}) : \varepsilon \in (-\delta, \delta)\}$ is also a parametric submodel of $\mathcal{P}_i$. Following the proof of Lemma 1, we have $s \in \mathcal{T}_i$, which is the tangent space of $\mathcal{P}_i$. This is true for all $i = 1, \dots, K$, so $s \in \mathcal{T}_1 \cap \dots \cap \mathcal{T}_K$ and therefore $\mathcal{T} \subseteq \mathcal{T}_1 \cap \dots \cap \mathcal{T}_K$.

For the other direction, pick any $f \in \mathcal{T}_1 \cap \dots \cap \mathcal{T}_K$. Define a parametric model by $p_\varepsilon(\mathbf{v}) = p_0(\mathbf{v})(1 + \varepsilon f(\mathbf{v}))$. Since $\mathbb{E}[f] = 0$, p_ε sums to 1. Moreover, the parameter $\varepsilon \in (-\delta, \delta)$ and δ is small enough so that $p_\varepsilon(\mathbf{v}) \geq 0$ at all $\mathbf{v}$. Then p_ε is a valid probability distribution. This parametric model contains p_0 at $\varepsilon = 0$, and the score function at p_0 is $\frac{\partial}{\partial \varepsilon} \log p_\varepsilon(\mathbf{v})|_{\varepsilon=0} = f$. What left to show is that all K constraints hold for all $p_\varepsilon(\mathbf{v})$. Indeed, for each i , $f \in \mathcal{T}_i$. Following the proof of Lemma 1, the i -th constraint $\mathbf{X}_i \perp\!\!\!\perp \mathbf{Y}_i \mid \mathbf{Z}_i$ must hold in $p_\varepsilon(\mathbf{v})$. Therefore, $\mathcal{T}_1 \cap \dots \cap \mathcal{T}_K \subseteq \mathcal{T}$. By definition, $\mathcal{T}$ is closed, so $\mathcal{T}_1 \cap \dots \cap \mathcal{T}_K \subseteq \mathcal{T}$.

We have shown that $\mathcal{T} = \overline{\mathcal{T}_1 \cap \dots \cap \mathcal{T}_K}$. Since $\mathcal{T}_1, \dots, \mathcal{T}_K$ are closed subspaces, $\mathcal{T} = \mathcal{T}_1 \cap \dots \cap \mathcal{T}_K$.

Next, we will show $\mathcal{T}^\perp = \mathcal{T}_1^\perp \oplus \dots \oplus \mathcal{T}_K^\perp$.

We have $\mathcal{T} = \mathcal{T}_1 \cap \dots \cap \mathcal{T}_K$ is orthogonal to all $\{\mathcal{T}_i^\perp\}_{i=1}^K$, so $\mathcal{T}$ is orthogonal to the span $\overline{\mathcal{T}_1^\perp \oplus \dots \oplus \mathcal{T}_K^\perp}$. Since $\mathcal{T}_1, \dots, \mathcal{T}_K$ are closed, this span is exactly the direct sum $\mathcal{T}_1^\perp \oplus \dots \oplus \mathcal{T}_K^\perp$. Therefore, $\mathcal{T}_1^\perp \oplus \dots \oplus \mathcal{T}_K^\perp \subseteq \mathcal{T}^\perp$.

For the other direction, every $f \in (\mathcal{T}_1^\perp \oplus \dots \oplus \mathcal{T}_K^\perp)^\perp$ must be orthogonal to all $\{\mathcal{T}_i^\perp\}_{i=1}^K$. For each i , the fact that f is orthogonal to $\mathcal{T}_i^\perp$ implies $f \in \mathcal{T}_i$. Therefore, $f \in \mathcal{T}_1 \cap \dots \cap \mathcal{T}_K = \mathcal{T}$. Hence, $(\mathcal{T}_1^\perp \oplus \dots \oplus \mathcal{T}_K^\perp)^\perp \subseteq \mathcal{T}$, and $\mathcal{T}_1^\perp \oplus \dots \oplus \mathcal{T}_K^\perp \supseteq \mathcal{T}^\perp$.

Finally, by Lemma 1, the i -th subspace $\mathcal{T}_i^\perp = \{\Pi(h_i \mid \mathcal{T}_i^\perp) : \forall h_i \in \mathcal{H}\}$. Then by definition of direct sum, the tangent space $\mathcal{T}$ is the subspace of all functions of the form $\sum_{i=1}^K \Pi(h_i \mid \mathcal{T}_i^\perp)$ for arbitrary $h_1, \dots, h_K \in \mathcal{H}$. □

Example 1. The Bell scenario model associated with Figure 1d has two constraints: $A \perp\!\!\!\perp B, D$ and $B \perp\!\!\!\perp A, C$. The operators correspond to these constraints are

$$\begin{aligned} \Pi(\cdot \mid \mathcal{T}_1^\perp) &= \mathbb{E}[\cdot | a, b, d] - \mathbb{E}[\cdot | a] - \mathbb{E}[\cdot | b, d] \\ \Pi(\cdot \mid \mathcal{T}_2^\perp) &= \mathbb{E}[\cdot | b, a, c] - \mathbb{E}[\cdot | b] - \mathbb{E}[\cdot | a, c]. \end{aligned} \quad (80)$$

Let $\Gamma(\cdot) = \Pi(\cdot \mid \mathcal{T}_1^\perp) + \Pi(\cdot \mid \mathcal{T}_2^\perp)$. If Γ is a projection operator, then $h - \Gamma(h)$ and $\Gamma(g)$ must be orthogonal, for any $h, g \in \mathcal{H}$. We will compute their inner product. It is not evident that their inner product vanishes.

$$\mathbb{E}[(h - \Gamma(h))\Gamma(g)] = \mathbb{E}[h\Gamma(g)] - \mathbb{E}[\Gamma(h)\Gamma(g)]$$

For the first term

$$\begin{aligned} \text{term1} &= \mathbb{E} \left[h \times \left(\mathbb{E}[g|A, B, D] - \mathbb{E}[g|A] - \mathbb{E}[g|B, D] + \mathbb{E}[g|B, A, C] - \mathbb{E}[g|B] - \mathbb{E}[g|A, C] \right) \right] \\ &= \mathbb{E}[\mathbb{E}[h|A, B, D] \cdot \mathbb{E}[g|A, B, D]] + \mathbb{E}[\mathbb{E}[h|A] \cdot \mathbb{E}[g|A]] - \mathbb{E}[\mathbb{E}[h|B, D] \cdot \mathbb{E}[g|B, D]] \\ &\quad + \mathbb{E}[\mathbb{E}[h|B, A, C] \cdot \mathbb{E}[g|B, A, C]] - \mathbb{E}[\mathbb{E}[h|B] \cdot \mathbb{E}[g|B]] - \mathbb{E}[\mathbb{E}[h|A, C] \cdot \mathbb{E}[g|A, C]] \end{aligned}$$

For the second term

$$\begin{aligned}
\text{term2} &= \mathbb{E} \left[\left(\mathbb{E}[h|A, B, D] - \mathbb{E}[h|A] - \mathbb{E}[h|B, D] + \mathbb{E}[h|B, A, C] - \mathbb{E}[h|B] - \mathbb{E}[h|A, C] \right) \right. \\
&\quad \times \left. \left(\mathbb{E}[g|A, B, D] - \mathbb{E}[g|A] - \mathbb{E}[g|B, D] + \mathbb{E}[g|B, A, C] - \mathbb{E}[g|B] - \mathbb{E}[g|A, C] \right) \right] \\
&= \mathbb{E} [\mathbb{E}[h|A, B, D] \cdot \mathbb{E}[g|A, B, D]] + \mathbb{E} [\mathbb{E}[h|A] \cdot \mathbb{E}[g|A]] - \mathbb{E} [\mathbb{E}[h|B, D] \cdot \mathbb{E}[g|B, D]] \quad (\text{term2a}) \\
&\quad + \mathbb{E} [\mathbb{E}[h|A, B, D] (\mathbb{E}[g|B, A, C] - \mathbb{E}[g|B] - \mathbb{E}[g|A, C])] \quad (\text{term2b}) \\
&\quad - \mathbb{E} [\mathbb{E}[h|A] (\mathbb{E}[g|A, B, D] - \mathbb{E}[g|A] - \mathbb{E}[g|B, D])] - \mathbb{E} [\mathbb{E}[h|A] (\mathbb{E}[g|B, A, C] - \mathbb{E}[g|B] - \mathbb{E}[g|A, C])] \quad (\text{term2c}) \\
&\quad - \mathbb{E} [\mathbb{E}[h|B, D] (\mathbb{E}[g|A, B, D] - \mathbb{E}[g|A] - \mathbb{E}[g|B, D])] - \mathbb{E} [\mathbb{E}[h|B, D] (\mathbb{E}[g|B, A, C] - \mathbb{E}[g|B] - \mathbb{E}[g|A, C])] \quad (\text{term2d}) \\
&\quad + \mathbb{E} [\mathbb{E}[h|B, A, C] \cdot \mathbb{E}[g|B, A, C]] - \mathbb{E} [\mathbb{E}[h|B] \cdot \mathbb{E}[g|B]] - \mathbb{E} [\mathbb{E}[h|A, C] \cdot \mathbb{E}[g|A, C]] \quad (\text{term2e}) \\
&\quad + \mathbb{E} [\mathbb{E}[h|B, A, C] (\mathbb{E}[g|A, B, D] - \mathbb{E}[g|A] - \mathbb{E}[g|B, D])] \quad (\text{term2f}) \\
&\quad - \mathbb{E} [\mathbb{E}[h|B] (\mathbb{E}[g|A, B, D] - \mathbb{E}[g|A] - \mathbb{E}[g|B, D])] - \mathbb{E} [\mathbb{E}[h|B] (\mathbb{E}[g|B, A, C] - \mathbb{E}[g|B] - \mathbb{E}[g|A, C])] \quad (\text{term2g}) \\
&\quad - \mathbb{E} [\mathbb{E}[h|A, C] (\mathbb{E}[g|A, B, D] - \mathbb{E}[g|A] - \mathbb{E}[g|B, D])] - \mathbb{E} [\mathbb{E}[h|A, C] (\mathbb{E}[g|B, A, C] - \mathbb{E}[g|B] - \mathbb{E}[g|A, C])] \quad (\text{term2h})
\end{aligned}$$

Note that $\text{term1} = \text{term2a} + \text{term2e}$. For the others,

$$\begin{aligned}
\text{term2b} &= \mathbb{E} [\mathbb{E}[h|A, B, D] \mathbb{E}[g|B, A, C]] - \mathbb{E} [\mathbb{E}[h|B] \mathbb{E}[g|B]] - \mathbb{E} [\mathbb{E}[h|A, B, D] \mathbb{E}[g|A, C]] \\
\text{term2c} &= -\cancel{\mathbb{E} [\mathbb{E}[h|A] \mathbb{E}[g|A]]} + \cancel{\mathbb{E} [\mathbb{E}[h|A] \mathbb{E}[g|A]]} + \mathbb{E} [\mathbb{E}[h|A] \mathbb{E}[g|B, D]] \\
&\quad - \cancel{\mathbb{E} [\mathbb{E}[h|A] \mathbb{E}[g|A]]} + \mathbb{E} [\mathbb{E}[h|A] \mathbb{E}[g|B]] + \cancel{\mathbb{E} [\mathbb{E}[h|A] \mathbb{E}[g|A]]} \\
\text{term2d} &= -\cancel{\mathbb{E} [\mathbb{E}[h|B, D] \mathbb{E}[g|B, D]]} + \mathbb{E} [\mathbb{E}[h|B, D] \mathbb{E}[g|A]] + \cancel{\mathbb{E} [\mathbb{E}[h|B, D] \mathbb{E}[g|B, D]]} \\
&\quad - \mathbb{E} [\mathbb{E}[h|B, D] \mathbb{E}[g|B, A, C]] + \mathbb{E} [\mathbb{E}[h|B] \mathbb{E}[g|B]] + \mathbb{E} [\mathbb{E}[h|B, D] \mathbb{E}[g|A, C]]
\end{aligned}$$

and

$$\begin{aligned}
\text{term2f} &= \mathbb{E} [\mathbb{E}[h|B, A, C] \mathbb{E}[g|A, B, D]] - \mathbb{E} [\mathbb{E}[h|A] \mathbb{E}[g|A]] - \mathbb{E} [\mathbb{E}[h|B, A, C] \mathbb{E}[g|B, D]] \\
\text{term2g} &= -\cancel{\mathbb{E} [\mathbb{E}[h|B] \mathbb{E}[g|B]]} + \mathbb{E} [\mathbb{E}[h|B] \mathbb{E}[g|A]] + \cancel{\mathbb{E} [\mathbb{E}[h|B] \mathbb{E}[g|B]]} \\
&\quad - \cancel{\mathbb{E} [\mathbb{E}[h|B] \mathbb{E}[g|B]]} + \cancel{\mathbb{E} [\mathbb{E}[h|B] \mathbb{E}[g|B]]} + \mathbb{E} [\mathbb{E}[h|B] \mathbb{E}[g|A, C]] \\
\text{term2h} &= -\mathbb{E} [\mathbb{E}[h|A, C] \mathbb{E}[g|A, B, D]] + \mathbb{E} [\mathbb{E}[h|A] \mathbb{E}[g|A]] + \mathbb{E} [\mathbb{E}[h|A, C] \mathbb{E}[g|B, D]] \\
&\quad - \cancel{\mathbb{E} [\mathbb{E}[h|A, C] \mathbb{E}[g|A, C]]} + \mathbb{E} [\mathbb{E}[h|A, C] \mathbb{E}[g|B]] + \cancel{\mathbb{E} [\mathbb{E}[h|A, C] \mathbb{E}[g|A, C]]}
\end{aligned}$$

Then

$$\text{term2} = \text{term1}$$

$$\begin{aligned} & + \mathbb{E} \left[\mathbb{E}[h|A, B, D] \mathbb{E}[g|B, A, C] \right] - \cancel{\mathbb{E} \left[\mathbb{E}[h|B] \mathbb{E}[g|B] \right]} - \mathbb{E} \left[\mathbb{E}[h|A, B, D] \mathbb{E}[g|A, C] \right] \\ & + \mathbb{E} \left[\mathbb{E}[h|A] \mathbb{E}[g|B, D] \right] + \mathbb{E} \left[\mathbb{E}[h|A] \mathbb{E}[g|B] \right] \\ & + \mathbb{E} \left[\mathbb{E}[h|B, D] \mathbb{E}[g|A] \right] - \mathbb{E} \left[\mathbb{E}[h|B, D] \mathbb{E}[g|B, A, C] \right] + \cancel{\mathbb{E} \left[\mathbb{E}[h|B] \mathbb{E}[g|B] \right]} + \mathbb{E} \left[\mathbb{E}[h|B, D] \mathbb{E}[g|A, C] \right] \\ & + \mathbb{E} \left[\mathbb{E}[h|B, A, C] \mathbb{E}[g|A, B, D] \right] - \cancel{\mathbb{E} \left[\mathbb{E}[h|A] \mathbb{E}[g|A] \right]} - \mathbb{E} \left[\mathbb{E}[h|B, A, C] \mathbb{E}[g|B, D] \right] \\ & + \mathbb{E} \left[\mathbb{E}[h|B] \mathbb{E}[g|A] \right] + \mathbb{E} \left[\mathbb{E}[h|B] \mathbb{E}[g|A, C] \right] \\ & - \mathbb{E} \left[\mathbb{E}[h|A, C] \mathbb{E}[g|A, B, D] \right] + \cancel{\mathbb{E} \left[\mathbb{E}[h|A] \mathbb{E}[g|A] \right]} + \mathbb{E} \left[\mathbb{E}[h|A, C] \mathbb{E}[g|B, D] \right] + \mathbb{E} \left[\mathbb{E}[h|A, C] \mathbb{E}[g|B] \right] \end{aligned}$$

It is not evident that $\text{term2} = \text{term1}$.

D PROOFS IN SECTION 5

Theorem 4. Consider the Markov model with $K \geq 1$ constraint in Theorem 3, and let φ_0 be an influence function of the target parameter $\psi(p)$. For any sequence of integers $i_1, i_2, \dots, i_M$ where each $i_m \in \{1, \dots, K\}$ for all $m \in \{1, \dots, M\}$, define

$$\varphi_m := \varphi_{m-1} - \Pi(\varphi_{m-1} \mid \mathcal{T}_{i_m}^\perp), \quad \forall m \in \{1, \dots, M\}. \quad (81)$$

In words, φ_m is the projection of φ_{m-1} onto the tangent space $\mathcal{T}_{i_m}$ of the DAG model corresponding to the i_m -th constraint. Then $\varphi_0, \dots, \varphi_M$ is a sequence of influence functions of $\psi(p)$ with non-increasing variance, meaning $E[\varphi_m^2] \leq E[\varphi_{m-1}^2]$ for all $m \in \{1, \dots, M\}$.

Proof. The proof of this theorem follows from the discussions in Section 5.2 of the main paper and Section A of the Appendix. $\square$

The iterative process described in Theorem 4 along with the explicit form of the projections $\Pi(\cdot \mid \mathcal{T}_{i_m}^\perp)$ in Theorem 3 suggests a natural procedure for constructing estimators. For example, given an IF of the following form (obtained from one step of the iteration):

$$\begin{aligned} \varphi_m &:= \varphi_{m-1} - \Pi(\varphi_{m-1} \mid \mathcal{T}_{i_m}^\perp) \\ &= \varphi_{m-1} - E[\varphi_{m-1} \mid \mathbf{x}_{i_m}, \mathbf{y}_{i_m}, \mathbf{z}_{i_m}] + E[\varphi_{m-1} \mid \mathbf{x}_{i_m}, \mathbf{z}_{i_m}] + E[\varphi_{m-1} \mid \mathbf{y}_{i_m}, \mathbf{z}_{i_m}] - E[\varphi_{m-1} \mid \mathbf{z}_{i_m}], \end{aligned} \quad (82)$$

an estimator may be constructed by adopting additional nuisance models for the four additional expectation terms. For some specific IFs, we maybe able to simplify φ_m because some terms from φ_{m-1} and the expectations may cancel.

Note that this is only a single step in the iterative process, where multiple operators are applied sequentially. If we want to use an M -th improvement of some initial IF φ_0 , the expression will involve nested expectations, and may become challenging to evaluate in practice. However, in some cases, we may derive explicit estimators that demonstrate efficiency improvement, as we demonstrate via the following example.

Example 2. Consider the Bell scenario model corresponding to Figure 1d, whose constraints are $A \perp\!\!\!\perp B, D$ and $B \perp\!\!\!\perp A, C$. Suppose that we want to estimate the target parameter $\psi(p) = \mathbb{E}[C \mid a_0]$ instead of $\mathbb{E}[D \mid a_0]$, as D and A are independent in this model. The influence function for this parameter in the saturated model is:

$$\varphi_0(a, c) = \frac{\mathbb{I}(a = a_0)}{p(a)}(c - E[C \mid a]) \quad (83)$$

Let $\mathcal{P}_1, \mathcal{P}_2$ be the single independence DAG models corresponding to the constraints $A \perp\!\!\!\perp B, D$ and $B \perp\!\!\!\perp A, C$, respectively. According to Theorem 3

$$\begin{aligned}\Pi(\cdot \mid \mathcal{T}_1^\perp) &= \mathbb{E}[\cdot \mid a, b, d] - \mathbb{E}[\cdot \mid a] - \mathbb{E}[\cdot \mid b, d] \\ \Pi(\cdot \mid \mathcal{T}_2^\perp) &= \mathbb{E}[\cdot \mid b, a, c] - \mathbb{E}[\cdot \mid b] - \mathbb{E}[\cdot \mid a, c].\end{aligned}$$

We will derive the first and second improvements of φ_0 , namely φ_1 and φ_2 , respectively, using the iterative method in Theorem 4, and explicitly derive the corresponding estimators.

First improvement:

$$\varphi_1 = \varphi_0 - \mathbb{E}[\varphi_0 \mid a, b, d] + \mathbb{E}[\varphi_0 \mid a] + \mathbb{E}[\varphi_0 \mid b, d], \quad (84)$$

where

$$\begin{aligned}\mathbb{E}[\varphi_0 \mid a, b, d] &= \frac{\mathbb{I}(a = a_0)}{p(a)} (\mathbb{E}[C \mid a, b, d] - \mathbb{E}[C \mid a]) \\ \mathbb{E}[\varphi_0 \mid a] &= \frac{\mathbb{I}(a = a_0)}{p(a)} (\mathbb{E}[C \mid a] - \mathbb{E}[C \mid a]) = 0 \\ \mathbb{E}[\varphi_0 \mid b, d] &= \int \mathbb{E}[\varphi_0 \mid a', b, d] p(a' \mid b, d) da' \\ &= \int \mathbb{E}[\varphi_0 \mid a', b, d] p(a') da' \quad (A \perp\!\!\!\perp B, D) \\ &= \int \left(\frac{\mathbb{I}(a' = a_0)}{p(a')} (\mathbb{E}[C \mid a', b, d] - \mathbb{E}[C \mid a']) \right) p(a') da' \\ &= \mathbb{E}[C \mid a_0, b, d] - \mathbb{E}[C \mid a_0]\end{aligned} \quad (85)$$

Substituting these expressions, cancelling terms, and noting that $\psi(p) = \mathbb{E}[C \mid a_0]$, we get

$$\varphi_1(a, b, c, d) = \frac{\mathbb{I}(a = a_0)}{p(a)} (c - \mathbb{E}[C \mid a, b, d]) + \mathbb{E}[C \mid a_0, b, d] - \psi(p) \quad (86)$$

Let $\hat{p}(a)$ be an estimate of the distribution $p(a)$, and $\hat{\mathbb{E}}[C \mid a, b, d]$ be an estimate of the conditional mean. The estimation equation method gives us an estimate $\hat{\psi}_1$ of the target $\psi(p)$ via

$$\hat{\psi}_1 = \frac{1}{n} \sum_{i=1}^n \left(\frac{\mathbb{I}(A_i = a_0)}{\hat{p}(A_i)} (C_i - \hat{\mathbb{E}}[C \mid A_i, B_i, D_i]) + \hat{\mathbb{E}}[C \mid a_0, B_i, D_i] \right). \quad (87)$$

Note that this estimator is the augmented inverse probability weighted (AIPW) estimator arising in causal inference and missing data problems. For example, A is the treatment, C is the outcome, and $\{B, D\}$ is the observed confounder. Then the Bell scenario is the observed confounder scenario, with randomization on A , explaining this “coincidence”.

Second improvement:

$$\varphi_2 = \varphi_1 - E[\varphi_1 \mid b, a, c] + E[\varphi_1 \mid b] + E[\varphi_1 \mid a, c], \quad (88)$$

where

$$\begin{aligned}\mathbb{E}[\varphi_1 \mid b] &= \int \frac{\mathbb{I}(a' = a_0)}{p(a')} c' \cdot p(a', c' \mid b) da' dc' - \int \frac{\mathbb{I}(a' = a_0)}{p(a')} \left(\int \mathbb{E}[C \mid a', b, d'] p(d' \mid b, a') dd' \right) p(a' \mid b) da' \\ &\quad + \int \mathbb{E}[C \mid a_0, b, d'] p(d' \mid b) dd' - \psi(p) \\ &= \mathbb{E}[C \mid a_0] - \int \mathbb{E}[C \mid a_0, b, d'] p(d' \mid b, a_0) dd' \quad (A, C \perp\!\!\!\perp B) \\ &\quad + \int \mathbb{E}[C \mid a_0, b, d'] p(d' \mid b) dd' - \psi(p) \\ &= \int \mathbb{E}[C \mid a_0, b, d'] p(d' \mid b) dd' - \int \mathbb{E}[C \mid a_0, b, d'] p(d' \mid b, a_0) dd' \quad (\mathbb{E}[C \mid a_0] = \psi(p)) \\ &= \int \mathbb{E}[C \mid a_0, b, d'] p(d' \mid b, a_0) dd' - \int \mathbb{E}[C \mid a_0, b, d'] p(d' \mid b, a_0) dd' \quad (D \perp\!\!\!\perp A \mid B) \\ &= 0.\end{aligned} \quad (89)$$

$$\begin{aligned}
\mathbb{E}[\varphi_1 \mid b, a, c] &= \frac{\mathbb{I}(a = a_0)}{p(a)} \left(c - \int \mathbb{E}[C \mid a, b, d'] p(d' \mid b, a, c) dd' \right) \\
&\quad + \int \mathbb{E}[C \mid a_0, b, d'] p(d' \mid b, a, c) dd' - \psi(p) \\
\mathbb{E}[\varphi_1 \mid a, c] &= \frac{\mathbb{I}(a = a_0)}{p(a)} \left(c - \int \mathbb{E}[C \mid a, b', d'] p(d' \mid b', a, c) p(b') dd' db' \right) \quad (B \perp\!\!\!\perp A, C) \\
&\quad + \int \mathbb{E}[C \mid a_0, b', d'] p(d' \mid b', a, c) p(b') dd' db' - \psi(p) \quad (B \perp\!\!\!\perp A, C).
\end{aligned} \tag{90}$$

Then

$$\begin{aligned}
\varphi_2(a, b, c, d) &= \frac{\mathbb{I}(a = a_0)}{p(a)} (c - \mathbb{E}[C \mid a, b, d]) + \mathbb{E}[C \mid a_0, b, d] - \psi(p) \\
&\quad + \frac{\mathbb{I}(a = a_0)}{p(a)} \left(\int \mathbb{E}[C \mid a, b, d'] p(d' \mid b, a, c) dd' - \int \mathbb{E}[C \mid a, b, d'] p(d' \mid b', a, c) p(b') dd' db' \right) \\
&\quad - \int \mathbb{E}[C \mid a_0, b, d'] p(d' \mid b, a, c) dd' + \int \mathbb{E}[C \mid a_0, b, d'] p(d' \mid b', a, c) p(b') dd' db'
\end{aligned} \tag{91}$$

We posit models for $\hat{p}(a)$, $\hat{\mathbb{E}}[C \mid a, b, d]$, $\hat{p}(d \mid a, c, b)$ and $\hat{p}(b)$. Importantly, these models must be consistent, i.e., there is a valid set of probability distributions $p(a, b, c, d)$ satisfying the constraints and yield these models. Enforcing consistency is not always obvious. If the models are consistent, then the estimation equation method gives us the estimator $\hat{\psi}_2$ for the target $\psi(p)$

$$\begin{aligned}
\hat{\psi}_2 &= \frac{1}{n} \sum_{i=1}^n \left(\frac{\mathbb{I}(A_i = a_0)}{\hat{p}(A_i)} (C_i - \hat{\mathbb{E}}[C \mid A_i, B_i, D_i]) + \hat{\mathbb{E}}[C \mid a_0, B_i, D_i] \right) \\
&\quad + \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}(A_i = a)}{\hat{p}(A_i)} \left(\int \hat{\mathbb{E}}[C \mid A_i, B_i, d'] \hat{p}(d' \mid B_i, A_i, C_i) dd' - \int \hat{\mathbb{E}}[C \mid A_i, b', d'] \hat{p}(d' \mid b', A_i, C_i) \hat{p}(b') dd' db' \right) \\
&\quad + \frac{1}{n} \sum_{i=1}^n \left(- \int \hat{\mathbb{E}}[C \mid a_0, B_i, d'] \hat{p}(d' \mid B_i, A_i, C_i) dd' + \int \hat{\mathbb{E}}[C \mid a_0, b', d'] \hat{p}(d' \mid b', A_i, C_i) \hat{p}(b') dd' db' \right).
\end{aligned} \tag{92}$$