Journal of Machine Learning Research 21 (2020) 1-86 Submitted 12/19; Revised 6/20; Published 9/20

Efficient Adjustment Sets for Population Average Causal Treatment Effect Estimation in Graphical Models

Andrea Rotnitzky AROTNITZKY@UTDT.EDU
Department of Economics
Universidad Torcuato Di Tella
Buenos Aires 1428, Argentina
and CONICET, Argentina

Ezequiel Smucler ESMUCLER@UTDT.EDU
Department of Mathematics and Statistics
Universidad Torcuato Di Tella
Buenos Aires 1428, Argentina

Editor: David Sontag

Abstract

The method of covariate adjustment is often used for estimation of total treatment effects from observational studies. Restricting attention to causal linear models, a recent article (Henckel et al., 2019) derived two novel graphical criteria: one to compare the asymptotic variance of linear regression treatment effect estimators that control for certain distinct adjustment sets and another to identify the optimal adjustment set that yields the least squares estimator with the smallest asymptotic variance. In this paper we show that the same graphical criteria can be used in non-parametric causal graphical models when treatment effects are estimated using non-parametrically adjusted estimators of the interventional means. We also provide a new graphical criterion for determining the optimal adjustment set among the minimal adjustment sets and another novel graphical criterion for comparing time dependent adjustment sets. We show that uniformly optimal time dependent adjustment sets do not always exist. For point interventions, we provide a sound and complete graphical criterion for determining when a non-parametric optimally adjusted estimator of an interventional mean, or of a contrast of interventional means, is semiparametric efficient under the non-parametric causal graphical model. In addition, when the criterion is not met, we provide a sound algorithm that checks for possible simplifications of the efficient influence function of the parameter. Finally, we find an interesting connection between identification and efficient covariate adjustment estimation. Specifically, we show that if there exists an identifying formula for an interventional mean that depends only on treatment, outcome and mediators, then the non-parametric optimally adjusted estimator can never be globally efficient under the causal graphical model.

Keywords: adjustment sets, back-door formula, Bayesian networks, causal inference, semiparametric inference

1. Introduction

Estimating total, population average, causal treatment effects by controlling for, that is, conditioning on, a subset of covariates is known as the method of covariate adjustment. Assuming a causal directed acyclic graph (DAG) model, the back-door criterion (Pearl,

2000) is a popular graphical criterion that gives sufficient conditions for a covariate set to be such that control for this set yields consistent estimators of total treatment effects. Shpitser et al. (2010) gives a necessary and sufficient graphical criterion for a subset of covariates to qualify for adjustment.

The graphical criteria of Pearl and Shpitser et al. are particularly useful for designing observational studies. Specifically, investigators planning an observational study might be prepared to hypothesize a causal diagram and apply the aforementioned criteria to aid them in selecting the covariates to measure in order to control for confounding. When many covariate adjustment sets are available, a natural question is which one should be selected.

Henckel et al. (2019) (see also Witte et al. (2020)) gave an answer to this question under the following assumptions: (i) the causal DAG model is linear, that is, each vertex in the DAG stands for a random variable that follows a linear regression model on its parents in the DAG, with an independent error that has an arbitrary distribution and (ii) the total treatment effects are estimated with the coefficients associated with treatments in the ordinary least squares (OLS) fit of the outcome on treatments and a set of valid adjustment covariates. They derive a graphical criterion that identifies the optimal covariate adjustment set in the sense that this set yields the OLS treatment effect estimator which has the smallest asymptotic variance among all OLS estimators of treatment effects that control for valid adjustment sets.

Our first contribution, see Section 5.1, is to establish that the same criterion holds for identifying the optimal valid covariate adjustment set when (i) the causal DAG model is non-parametric in the sense that no assumptions are made on the conditional distribution of each node given is parents, and, (ii) the treatment effects are estimated non-parametrically, that is, without exploiting the conditional independencies in the data generating law encoded in the causal DAG model. For instance, the treatment effects could be estimated by inverse probability weighting with the propensity score estimated non-parametrically (Hirano et al., 2003; Abadie and Cattaneo, 2018), or by doubly-robust or double-machine learning approaches (Chernozhukov et al.; Smucler et al., 2019). Our second contribution is to provide a graphical criterion for identifying the optimal adjustment set among the class of minimal adjustment sets. A minimal adjustment set is a valid adjustment set such that removal of any vertex from the set yields a non-valid adjustment set. We note that our criterion holds for non-parametric causal DAG models and estimators as well as linear causal DAG models and estimators.

A second important contribution of Henckel et al. (2019) is a graphical criterion, assuming linear DAG models and OLS estimators, to compare certain pairs of valid adjustment sets which is more broadly applicable than earlier existing criteria (Kuroki and Miyakawa, 2003; Kuroki and Cai, 2004). Building on their criterion Henckel et al. also provided a simple procedure that, for a valid adjustment set, returns a pruned valid adjustment set that yields OLS estimators of treatment effects with smaller asymptotic variance. The procedure was conjectured to yield improved efficiency in VanderWeele and Shpitser (2011). The contribution of Henckel et al. (2019) was to rigorously show that the conjecture is valid for causal linear models and OLS estimators of treatment effects. Our third contribution is to prove that both the graphical criterion and the pruning procedure of Henckel et al. (2019) also apply for non-parametric causal DAG models and estimators.

Henckel et al. (2019) considered not only DAGs but also (linear) completed partially directed acyclic graphs (CPDAGs) and maximal PDAGs. A CPDAG (Meek, 1995; Andersson et al., 1997; Spirtes et al., 2000; Chickering, 2002) represents, under causal sufficiency and faithfulness, the Markov equivalence class of DAGs that can be deduced from the conditional independencies in the observed data distribution. A maximal PDAGs is a maximally oriented partially directed acyclic graph that maximally refines the Markov equivalence class when the orientation of some edges are known a-priori (Meek, 1995; Scheines et al., 1998; Hoyer et al., 2008; Hauser and Bühlmann, 2012; Eigenmann et al., 2017; Wang et al., 2017). Henckel et al. (2019) derived graphical criteria for identifying the optimal adjustment set and for comparing certain adjustment sets under linear CPDAGs and maximal PDAGs, assuming treatment effects are estimated by least squares. These criteria are consequences of the corresponding criteria for DAGs. This is because the criteria are based solely on d-separation conditions on CPDAGs and maximal PDAGs, and d-separations that hold on CPDAGs and maximal PDAGs hold on all possible DAGs represented by them. Because, as indicated earlier, we show that the graphical criteria developed by Henckel et al. (2019) for linear DAGs and estimators also holds for non-parametric DAGs and estimators, we conclude that the criteria derived by Henckel et al. (2019) for linear CPDAGs and maximal PDAGs using linear estimators of treatment effects, also hold for non-parametric CPDAGs and maximal PDAGs when non-parametric estimators of treatment effects are used. To avoid repetitions we do not expand on this topic in the present paper and refer the reader to Henckel et al. (2019).

The aforementioned graphical criterion of Henckel et al. (2019) for comparing certain adjustment sets in DAGs applies to OLS estimators of the causal effects of both point and joint interventions. However, for joint interventions, the criterion makes the restrictive assumption that the adjustment sets are time independent. As Henckel et al. (2019) pointed out, time independent covariate adjustment sets for joint interventions do not always exist. In contrast, time *dependent* covariate adjustment sets, which are comprised by covariates that are needed to adjust for future treatments but are themselves affected by earlier treatments, always exist. The g-formula (Robins, 1986), is the generalization of the adjustment formula from time independent to time dependent covariate adjustment sets. This raises the question of whether it is possible to generalize the results obtained for comparing time independent covariate adjustment sets to time dependent covariate adjustment sets. The answer is mixed. Specifically, in Section 5.2 we establish a result (Theorem 13) that allows the comparison of certain time dependent covariate adjustment sets and which generalizes the results obtained for non-parametric models and estimators in Theorem 6 of the present article from time independent to time dependent covariate adjustment sets. However, in that section we also exhibit a DAG in which no uniformly optimal time dependent covariate adjustment set exists. We do so by exhibiting two data generating laws, both satisfying the restrictions implied by the non-parametric causal DAG, such that a given time dependent covariate adjustment set dominates all others for one law, in the sense of yielding non-parametric estimators of the g-formula with smallest asymptotic variance, but for the second law a different time dependent covariate adjustment set dominates the rest.

Next we investigate the following problem. If we could measure all the variables of the causal DAG, we could then exploit the conditional independencies encoded in the non-parametric causal DAG model to efficiently estimate the total treatment effects. For a point

exposure, we can also estimate each treatment effect by the method of covariate adjustment using the optimal time independent covariate adjustment set. A natural question then is under which DAG configurations, if any, do the two procedures result in estimators with the same asymptotic efficiency? From a practical perspective this question is interesting for the planning of observational studies since for DAGs for which no efficiency loss is incurred by non-parametric optimal covariate adjustment estimation, then the optimal covariate adjustment set, the treatment and the outcome are all the variables that one needs to measure not only for consistent but also for efficient estimation of treatment effects. In Section 6.1 we review a general one-step estimation strategy for computing semiparametric efficient estimators. We argue that only variables entering the efficient influence function of a interventional mean under the non-parametric causal graphical model are required for computing the one-step estimator of treatment effects. As such, all variables that do not enter into the efficient influence function are irrelevant for efficient estimation. In Section 6.2 we establish a sound and complete graphical criterion to determine whether or not the optimally adjusted non-parametric estimator incurs in loss of efficiency. The completeness of our criterion and of the ID algorithm (Tian and Pearl, 2002; Shpitser and Pearl, 2008) imply the following interesting result, established in Section 6.3, linking identification and efficient covariate adjustment estimation: if there exists an identifying formula for an interventional mean that depends only on treatment, outcome and mediators, then the non-parametric optimally adjusted estimator can never be globally efficient under the causal DAG model.

When the optimally adjusted estimator is not efficient, it may nevertheless be the case that not all the variables in the DAG enter into the calculation of an efficient estimator. As such, from the perspective of planning a study, it is useful to learn which variables are irrelevant for efficient estimation since such variables need not be measured. In Section 6.4 we provide a sound algorithm that checks for variables that do not enter into the efficient influence function and hence are irrelevant for efficient estimation. In addition, the algorithm conducts sound checks for possible simplifications of the formula for the efficient influence function.

The rest of the paper is organized as follows. In Section 2 we review some concepts of causal graphical models and semiparametric efficiency theory used throughout the paper. In Section 3 we review the definition of time independent adjustment sets and provide the definition of time dependent adjustment sets. In Section 4 we review the asymptotic theory of estimators based on the method of non-parametric covariate adjustment. In Section 5 we provide the main results concerning optimal adjustment sets. In Section 6 we discuss efficient estimation exploiting the restrictions of the causal graphical model. Section 7 concludes with a list of open problems. Proofs of all the results stated in the main text are given in the Appendix.

2. Background

In this section we review some elements of the theory of causal graphical models and of semiparametric efficiency theory that will be used throughout the paper.

2.1 Directed Acyclic Graphs

Directed graph. A directed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ consists of a finite node set $\mathbf{V}$ and a set of directed edges $\mathbf{E}$. A directed edge between two nodes V, W is represented by $V \to W$. Given a set of nodes $\mathbf{Z} \subset \mathbf{V}$ the induced subgraph $\mathcal{G}_{\mathbf{Z}} = (\mathbf{Z}, \mathbf{E}_Z)$ is the graph obtained by considering only nodes in $\mathbf{Z}$ and edges between nodes in $\mathbf{Z}$.
Paths. Two nodes are adjacent if there exists an edge between them. A path from a node V to a node W in graph $\mathcal{G}$ is a sequence of nodes $(V_1, \ldots, V_j)$ such that $V_1 = V$, $V_j = W$ and V_i and V_{i+1} are adjacent in $\mathcal{G}$ for all $i \in \{1, \ldots, j-1\}$. Then V and W are called the endpoints of the path. A path $(V_1, \ldots, V_j)$ is directed or causal if $V_i \to V_{i+1}$ for all $i \in \{1, \ldots, j-1\}$.
Ancestry. If $V \to W$, then V is a parent of W and W is a child of V. If there is a directed path from V to W, then V is an ancestor of W and W a descendant of V. We follow the convention that every node is an ancestor and a descendant of itself. The sets of parents, children, ancestors and descendants of V in $\mathcal{G}$ are denoted by $\mathrm{pa}_{\mathcal{G}}(V)$, $\mathrm{ch}_{\mathcal{G}}(V)$, $\mathrm{an}_{\mathcal{G}}(V)$, $\mathrm{de}_{\mathcal{G}}(V)$. The set of non-descendants of a vertex V is defined as $\mathrm{nd}_{\mathcal{G}}(V) \equiv \mathrm{de}^{\mathrm{c}}_{\mathcal{G}}(V)$.
Colliders and forks. A node V is a collider on a path δ if δ contains a subpath (U, V, W) such that $U \to V \leftarrow W$. A node V is called a fork on δ if δ contains a subpath (U, V, W) such that $U \leftarrow V \to W$.
Directed cycles, DAGs. A directed path from V to W, together with the edge $W \to V$ forms a directed cycle. A directed graph without directed cycles is called a directed acyclic graph (DAG). The nodes $(V_{k_1}, \ldots, V_{k_s})$ are said to follow a topological order relative to a DAG $\mathcal{G}$ if V_{k_j} is not an ancestor of $V_{k_{j'}}$ in $\mathcal{G}$ whenever $j > j'$.
d-separation (Pearl, 2000). Consider a DAG $\mathcal{G}$ and distinct sets of nodes $\mathbf{U}, \mathbf{W}, \mathbf{Z}$. A path δ between $U \in \mathbf{U}$ and $W \in \mathbf{W}$ is blocked by $\mathbf{Z}$ in $\mathcal{G}$ if one of the following holds:

1. δ contains a node that is not a collider and is a member of $\mathbf{Z}$, or
2. If there exists a collider C in δ such that neither C nor its descendants are in $\mathbf{Z}$.

$\mathbf{U}, \mathbf{W}$ are d-separated by $\mathbf{Z}$ in $\mathcal{G}$ (denoted as $\mathbf{U} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{W} \mid \mathbf{Z}$) if for any $U \in \mathbf{U}$ and $W \in \mathbf{W}$, all paths between U and W are blocked given $\mathbf{Z}$.
Bayesian Network. Given a DAG $\mathcal{G}$ with a vertex set $\mathbf{V}$ that represents a random vector defined on a given probability space, a law P for $\mathbf{V}$ is said to satisfy the Local Markov Property relative to $\mathcal{G}$ if and only if

$$V \perp\!\!\!\perp \mathrm{nd}_{\mathcal{G}}(V) \mid \mathrm{pa}_{\mathcal{G}}(V) \text{ under } P \text{ for all } V \in \mathbf{V},$$

where throughout if U and V are independent random variables defined on a common probability space we write $U \perp\!\!\!\perp V$.

The Bayesian Network represented by DAG $\mathcal{G}$ (Pearl, 2000) is defined as the collection

$$\mathcal{M}(\mathcal{G}) \equiv \{P : P \text{ satisfies the Local Markov Property relative to } \mathcal{G}\}.$$

Verma and Pearl (1990) and Geiger et al. (1990) show that for any disjoint sets $\mathbf{A}, \mathbf{B}, \mathbf{C}$ included in $\mathbf{V}$

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{C} \Leftrightarrow \mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C} \text{ under } P \text{ for all } P \in \mathcal{M}(\mathcal{G}).$$

Marginal DAG model (Evans, 2016). Let $\mathcal{G}$ be a DAG with vertices $\mathbf{V} \dot{\cup} \mathbf{U}$, where $\dot{}$ stands for disjoint union, and $\mathcal{V}$ a state-space for $\mathbf{V}$. Let $\mathcal{M}(\mathcal{G})$ be the Bayesian Network represented by $\mathcal{G}$. Define the *marginal DAG model* $\mathcal{M}(\mathcal{G}, \mathbf{V})$ by the collection of probability distributions P over $\mathbf{V}$ such that there exist

1. some state-space $\mathcal{U}$ for $\mathbf{U}$,
2. a probability measure $Q \in \mathcal{M}(\mathcal{G})$ over $\mathcal{V} \times \mathcal{U}$,

and P is the marginal distribution of Q over $\mathbf{V}$.
Exogenized DAG (Evans, 2016). Let $\mathcal{G}$ be a DAG and let U be a vertex of $\mathcal{G}$ with a single child R. Define the exogenized DAG $\tau(\mathcal{G}, U)$ as follows: take the vertices and edges of $\mathcal{G}$, and then (i) add an edge $H \to R$ from every $H \in \mathrm{pa}_{\mathcal{G}}(U)$ to R, and (ii) delete U and any edge $H \to U$ for $H \in \mathrm{pa}_{\mathcal{G}}(U)$. All other edges and vertices are as in $\mathcal{G}$. In words, to exogenize a DAG $\mathcal{G}$ relative to a vertex U with a single child, we join all parents of U to the child of U with directed edges, and then remove U and all edges into and out of U.
Latent projection. Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V} \cup \mathbf{L}$, where the the vertices in $\mathbf{V}$ are observable and the vertices in $\mathbf{L}$ are hidden. The latent projection (Verma and Pearl, 1990) $\mathcal{G}[\mathbf{V}]$ is a directed mixed graph (that is, a graph with both directed and bi-directed edges) with vertex set $\mathbf{V}$, where for each pair of distinct vertices $V_i, V_j \in \mathbf{V}$:

1. $\mathcal{G}[\mathbf{V}]$ contains $V_i \to V_j$ if and only if there exists a directed path from V_i to V_j on which every non-endpoint vertex is in $\mathbf{L}$.
2. $\mathcal{G}[\mathbf{V}]$ contains $V_i \leftrightarrow V_j$ if and only if there exists a path of the form $V_i \leftarrow \cdots \to V_j$, on which every non-endpoint vertex is a non-collider and an element of $\mathbf{L}$

District We call a set $\mathbf{D} \subset \mathbf{V}$ a bi-directed component of $\mathcal{G}[\mathbf{V}]$ if for any $U, W \in \mathbf{D}$ there exists a path from U to W in $\mathcal{G}[\mathbf{V}]$ of the form $U \leftrightarrow \cdots \leftrightarrow W$. $\mathbf{D} \subset \mathbf{V}$ is a district in $\mathcal{G}[\mathbf{V}]$ if it is an inclusion maximal bi-directed component.

Throughout we use standard set theory notation. For a DAG with node set $\mathbf{V}$ and for $\mathbf{U}, \mathbf{W} \subset \mathbf{V}$ we have $\mathbf{U}^c = \mathbf{V} \setminus \mathbf{U}$, $\mathbf{U} \setminus \mathbf{W} = \mathbf{U} \cap \mathbf{W}^c$ and $\mathbf{U} \triangle \mathbf{W} = (\mathbf{U} \setminus \mathbf{W}) \cup (\mathbf{W} \setminus \mathbf{U})$. For a vector $\mathbf{U} = (U_0, \ldots, U_r) \subset \mathbf{V}$ and $j \leq r$ we let

$$\overline{\mathbf{U}}_j \equiv (U_0, \ldots, U_j)\,.$$

2.2 Causal Graphical Models

A causal (agnostic) graphical model (Spirtes et al., 2000; Robins and Richardson, 2010) represented by $\mathcal{G}$ assumes that the law of $\mathbf{V} \equiv (V_1, \ldots, V_s)$ belongs to $\mathcal{M}(\mathcal{G})$ and that for any $\mathbf{A} = \{A_1, \ldots, A_p\} \subset \mathbf{V}$, the post-intervention density (with respect to a dominating measure) $f[\mathbf{v} \mid \mathrm{do}(\mathbf{a})]$ of $\mathbf{V}$ when $\mathbf{A}$ is set to $\mathbf{a}$ on the entire population satisfies

$$f[\mathbf{v} \mid \mathrm{do}(\mathbf{a})] = \begin{cases} \prod_{V_j \in \mathbf{V} \setminus \mathbf{A}} f(v_j \mid \mathrm{pa}_{\mathcal{G}}(V_j)) & \text{if } \mathbf{A} = \mathbf{a} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Formula (1) is known as the g-formula (Robins, 1986), the manipulated density formula (Spirtes et al., 2000) and the truncated factorization formula (Pearl, 2000).

The non-parametric structural equations model with independent errors (NPSEM-IE, Pearl 2000) is a sub-model of the causal agnostic graphical model that additionally assumes the existence of counterfactuals. Specifically, the model associates each vertex $V \in \mathbf{V}$ with a factual random variable satisfying

$$V = g_V\left(\mathrm{pa}_{\mathcal{G}}(V), \varepsilon_V\right) \text{ for all } V \in \mathbf{V}$$

where $\{\varepsilon_V\}_{V\in\mathbf{V}}$ are mutually independent and $\{g_V\}_{V\in\mathbf{V}}$ are arbitrary functions. The model also assumes that for any $\mathbf{A} = \{A_1, \ldots, A_p\} \subset \mathbf{V}$, the counterfactual vector $\mathbf{V_a}$ that would be observed had $\mathbf{A}$ been set to $\mathbf{a}$ exists, and is generated according to

$$\begin{aligned} V_{\mathbf{a}} &= g_V\left(\left[\mathrm{pa}_{\mathcal{G}}(V)\right]_{\mathbf{a}}, \varepsilon_V\right) \text{ for all } V \in \mathbf{V}\backslash\mathbf{A} \\ A_{\mathbf{a},k} &= a_k \qquad\qquad\qquad\qquad \text{for all } k = 1, \ldots, p. \end{aligned}$$

The finest fully randomized causally interpretable structured tree graph model (FFRCISTG, Robins 1986) makes the same assumptions as the NPSEM-IE model, except that it relaxes the assumption that the $\{\varepsilon_V\}_{V\in\mathbf{V}}$ are mutually independent. We note that the only restriction that the NPSEM-IE and the FFRCISTG models place on the law P of the factual random vector $\mathbf{V}$, is that $P \in \mathcal{M}(\mathcal{G})$. Furthermore, (1) remains valid under both models. See Richardson and Robins (2013) for more details.

The results that we will derive in this paper rely solely on the assumption that $P \in \mathcal{M}(\mathcal{G})$ and on the validity of (1). Therefore, the results hold for the causal agnostic graphical models, the NPSEM-IE, and the FFRCISTG.

A causal (agnostic) graphical linear model represented by $\mathcal{G}$ is the submodel of the causal (agnostic) graphical model which additionally imposes the restriction that $\mathbf{V} = (V_1, \ldots, V_s)$ satisfies

$$V_i = \sum_{V_j \in \mathrm{pa}_{\mathcal{G}}(V_i)} \alpha_{ij} V_j + \varepsilon_i,$$

for $i \in \{1, \ldots, S\}$, where $\alpha_{ij} \in \mathbb{R}$ and $\varepsilon_1, \ldots, \varepsilon_p$ are jointly independent random variables with zero mean and finite variance.

Throughout this paper we let $\mathbf{V_a}$ be a random vector with density $f[\mathbf{v} \mid \mathrm{do}(\mathbf{a})]$. In particular for $Y \in \mathbf{V}$ we let $Y_{\mathbf{a}}$ be the corresponding component of $\mathbf{V_a}$. We call $E[Y_{\mathbf{a}}] = E[Y \mid \mathrm{do}(\mathbf{a})]$ the interventional mean under $\mathbf{A} = \mathbf{a}$.

2.3 Semiparametric Efficiency Theory

We now review the key elements of semiparametric efficiency theory that we will use throughout the paper.

2.3.1 Asymptotically Linear Estimators

An estimator $\widehat{\gamma}$ of a scalar parameter $\gamma(P)$ based on n i.i.d. copies $\mathbf{V}_1, \ldots, \mathbf{V}_n$ of $\mathbf{V}$ is asymptotically linear at P if there exists a random variable $\varphi_P(\mathbf{V})$ with $E_P[\varphi_P(\mathbf{V})] = 0$ and $var[\varphi_P(\mathbf{V})] < +\infty$ such that under P

$$n^{1/2}\left\{\widehat{\gamma} - \gamma(P)\right\} = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \varphi_P(\mathbf{V}_i) + o_p(1).$$

Here and throughout $E_P[\cdot]$ and $var_P[\cdot]$ denote the mean and the variance operators under the law P. A random variable $\varphi_P(\mathbf{V})$ that satisfies the aforementioned conditions is unique almost surely P. It is called the influence function of $\widehat{\gamma}$ at P. By the Central Limit Theorem any asymptotically linear estimator $\widehat{\gamma}$ satisfies

$$n^{1/2}\{\widehat{\gamma} - \gamma(P)\} \xrightarrow{d} N(0, var_P[\varphi_P(\mathbf{V}_i)]),$$

where $\xrightarrow{d}$ is convergence in distribution under law P. Furthermore any two asymptotically linear estimators, say $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$, with the same influence function are asymptotically equivalent in the sense that $n^{1/2}(\widehat{\gamma}_1 - \widehat{\gamma}_2) = o_p(1)$.

2.3.2 Regular Estimators

Given a collection of probability laws $\mathcal{M}$ for $\mathbf{V}$, an estimator $\widehat{\gamma}$ of a scalar parameters $\gamma(P)$ is regular in $\mathcal{M}$ at P if its convergence to $\gamma(P)$ is locally uniform (Van der Vaart, 2000, Chapter 8, page 115). Regularity is a necessary condition for a nominal $1-\alpha$ level Wald interval centered at the estimator to be an honest confidence interval in the sense that there exists a sample size n^* such that for all $n > n^*$ the interval attains at least its nominal coverage over all laws in $\mathcal{M}$.

2.3.3 Efficiency

Given a collection of probability laws $\mathcal{M}$, for any law P in $\mathcal{M}$, define the tangent space $\Lambda \equiv \Lambda(P)$ at P of model $\mathcal{M}$ as the $L_2(P)-$closed linear span of scores at $t=0$ for regular one-dimensional parametric submodels $t \in [0,\varepsilon) \to P_t$ with $P_{t=0} = P$ (Van der Vaart, 2000, Chapter 25, page 362). If for all $P \in \mathcal{M}$, $\Lambda(P)$ is a subset of a euclidean space, the model $\mathcal{M}$ is said to be parametric. If for all $P \in \mathcal{M}$, $\Lambda(P)$ is a equal to $L_2(P)$, the model $\mathcal{M}$ is said to be non-parametric. Otherwise, the model is said to be semiparametric.

A parameter $\gamma(P)$, more precisely the map $P' \in \mathcal{M} \to \gamma(P')$, is pathwise differentiable at P if there exists a random variable $\xi_P(\mathbf{V})$ such that $E_P\left[\varphi_P(\mathbf{V};\mathcal{G})^2\right] < \infty$, $E_P[\xi_P(\mathbf{V})] = 0$ and such that for any regular one-dimensional parametric submodel $t \in [0,\varepsilon) \to P_t$ with $P_{t=0} = P$ and score at $t=0$ denoted as S, it holds that $d\gamma(P_t)/dt|_{t=0} = E_P[\xi_P(\mathbf{V})S]$. The random variable $\xi_P(\mathbf{V})$ is called an influence function of the parameter $\gamma(P)$. Unless $\mathcal{M}$ is non-parametric, there exists infinitely many influence functions, because if ξ_P is an influence function so is $\xi_P + T$ for any mean zero T uncorrelated with the elements of Λ.

The following key result in semiparametric theory connects the influence function of regular and asymptotically linear estimators with the influence functions of regular parameters. Specifically, if $\widehat{\gamma}$ is an asymptotically linear estimator of $\gamma(P)$ at P with influence function ξ_P, then $\widehat{\gamma}$ is regular at P in model $\mathcal{M}$ if and only if $\gamma(P)$ is pathwise differentiable at P and ξ_P is an influence function of $\gamma(P)$. See Theorem 2.2 of Newey (1990).

The projection $\Pi[B|\Lambda]$ of any $B \in L_2(P)$ into the tangent space Λ at P is defined as the unique element of Λ such that $B - \Pi[B|\Lambda]$ is uncorrelated under P with any element of Λ. The projection $\varphi_{P,eff} \equiv \Pi[\varphi_P(\mathbf{V})|\Lambda]$ of any influence function φ_P of $\gamma(P)$ is itself an influence function. $\varphi_{P,eff}$ is called *the efficient influence function* of $\gamma(P)$ at P in model $\mathcal{M}$. It follows from the Pythagorean Theorem, that the variance $\Omega_{eff} \equiv E_P\left[(\varphi_{P,eff})^2\right]$ of $\varphi_{P,eff}(\mathbf{V})$ is less than or equal to the variance $E_P\left[\varphi_P^2(\mathbf{V})\right]$ of any influence function

$\varphi_P(\mathbf{V})$. Consequently, Ω_{eff} is a lower bound for the variance of the limiting mean zero normal distribution of regular asymptotically linear estimators of $\gamma(P)$. Ω_{eff} is called the semiparametric variance bound (also called the semiparametric Cramer-Rao bound) for $\gamma(P)$ at P in model $\mathcal{M}$.

3. Interventional Mean and Adjustment Sets

Under the causal graphical model, for any $\mathbf{A} = \{A_0, \ldots, A_p\} \subset \mathbf{V}$ topologically ordered, where each A_k a discrete random variable and $Y \in \mathbf{V} \backslash \mathbf{A}$, the interventional mean on the outcome Y satisfies

$$E\left[Y_{\mathbf{a}}\right] = E_P\left[\prod_{k=0}^{p}\left\{\frac{I_{a_k}\left(A_k\right)}{P\left(A_k = a_k | \operatorname{pa}_{\mathcal{G}}\left(A_k\right)\right)}\right\} Y\right].$$

This is an immediate consequence of formula (1). The Local Markov Property for $P \in \mathcal{M}(\mathcal{G})$ further implies that $E\left[Y_{\mathbf{a}}\right]$ is equal to

$$E_P\left\{E_P\left\{E_P\left[E_P\left[Y|\mathbf{a}, \overline{\operatorname{pa}_{\mathcal{G}}(A_p)}\right]|\overline{\mathbf{a}}_{p-1}, \overline{\operatorname{pa}_{\mathcal{G}}(A_{p-1})}\right]|\overline{\mathbf{a}}_{p-2}, \overline{\operatorname{pa}_{\mathcal{G}}(A_{p-2})}\right\}\cdots | \mathbf{a}_0, \overline{\operatorname{pa}_{\mathcal{G}}(A_0)}\right\}$$

where for every $j \in \{0, \ldots, p\}$

$$\overline{\operatorname{pa}_{\mathcal{G}}(A_j)} = \bigcup_{k=0}^{j} \operatorname{pa}_{\mathcal{G}}(A_j).$$

See Robins (1987a), Theorem AD.1. In particular, if $\mathbf{A}$ is a point intervention, so that it is a single variable A, then

$$\begin{aligned} E\left[Y_a\right] &= E_P\left[\frac{I_a\left(A\right)}{P\left(A = a | \operatorname{pa}_{\mathcal{G}}\left(A\right)\right)} Y\right] \\ &= E_P\left[E\left[Y|A = a, \operatorname{pa}_{\mathcal{G}}\left(A\right)\right]\right]. \end{aligned} \tag{2}$$

For a binary point intervention A, the average treatment effect (ATE), $ATE \equiv E\left[Y_{a=1}\right] - E\left[Y_{a=0}\right]$, quantifies the effect on the mean of the outcome of setting $A = 1$ versus $A = 0$ on the entire population. Under a causal graphical model, equation (1) implies

$$ATE = E_P\left[E_P\left[Y|A = 1, \operatorname{pa}_{\mathcal{G}}\left(A\right)\right]\right] - E_P\left[E_P\left[Y|A = 0, \operatorname{pa}_{\mathcal{G}}\left(A\right)\right]\right].$$

3.1 Adjustment Sets

Definition 1 (Time dependent covariate adjustment set)
Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$ let $\mathbf{A} = (A_0, \ldots, A_p) \subset \mathbf{V}$ be topologically ordered and $Y \in \mathbf{V} \backslash \mathbf{A}$. We say that $\mathbf{Z} \equiv (\mathbf{Z}_0, \mathbf{Z}_1, \ldots, \mathbf{Z}_p) \subset \mathbf{V} \backslash \{\mathbf{A}, Y\}$ where $\mathbf{Z}_0, \mathbf{Z}_1, \ldots$ and $\mathbf{Z}_p$ are disjoint, is a time dependent covariate adjustment set relative to $(\mathbf{A}, Y)$ in $\mathcal{G}$ if under all $P \in \mathcal{M}\left(\mathcal{G}\right)$ and all $y \in \mathbb{R}$

$$\begin{aligned} &E_P\left[\prod_{k=0}^{p}\left\{\frac{I_{a_k}\left(A_k\right)}{P\left(A_k = a_k | \operatorname{pa}_{\mathcal{G}}\left(A_k\right)\right)}\right\} I_{(-\infty, y]}(Y)\right] = \\ &E_P\left\{E_P\left\{E_P\left[E_P\left[I_{(-\infty, y]}(Y)|\mathbf{a}, \mathbf{Z}\right]|\overline{\mathbf{a}}_{p-1}, \overline{\mathbf{Z}}_{p-1}\right]|\overline{\mathbf{a}}_{p-2}, \overline{\mathbf{Z}}_{p-2}\right\}\cdots | \mathbf{a}_0, \mathbf{Z}_0\right\}. \end{aligned}$$

The preceding definition extends the following definition of covariate adjustment set of Shpitser et al. (2010) and Maathuis and Colombo (2015). We use the appellatives time dependent and time independent to distinguish the two definitions.

Definition 2 (Time independent covariate adjustment set) *(Shpitser et al., 2010; Maathuis and Colombo, 2015) Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$ let $\mathbf{A} = (A_0, \ldots, A_p) \subset \mathbf{V}$ be topologically ordered and $Y \in \mathbf{V} \backslash \mathbf{A}$. A set $\mathbf{Z} \subset \mathbf{V} \backslash \{\mathbf{A}, Y\}$ is a time independent adjustment set relative to $(\mathbf{A}, Y)$ in $\mathcal{G}$ if under all $P \in \mathcal{M}(\mathcal{G})$ and all $y \in \mathbb{R}$*

$$E_P\left[\prod_{k=0}^{p}\left\{\frac{I_{a_k}(A_k)}{P\left(A_k = a_k | \operatorname{pa}_{\mathcal{G}}(A_k)\right)}\right\} I_{(-\infty,y]}(Y)\right] = E_P\left[E_P\left[I_{(-\infty,y]}(Y)|\mathbf{A}=\mathbf{a}, \mathbf{Z}\right]\right] \quad (3)$$

Note that $\mathbf{Z}$ is a time independent adjustment set if and only if $\widetilde{\mathbf{Z}} = (\mathbf{Z}_0, \ldots, \mathbf{Z}_p)$ with $\mathbf{Z}_0 = \mathbf{Z}$ and $\mathbf{Z}_j = \emptyset$ for $j = 1 \ldots, p$ is a time dependent adjustment set.

The back-door criterion (Pearl, 2000) is a sufficient graphical condition for $\mathbf{Z}$ to be a time independent adjustment set. Shpitser et al. (2010) gives a necessary and sufficient graphical condition for $\mathbf{Z}$ to be a time independent covariate adjustment set. These authors also show that if $\mathbf{Z}$ is a time independent covariate adjustment set, then there exists $\mathbf{Z}_{sub} \subset \mathbf{Z}$ such that $\mathbf{Z}_{sub}$ is a time independent adjustment set and it satisfies the back-door criterion. On the other hand Pearl and Robins (1995) provides a sufficient graphical criterion for $\mathbf{Z}$ to be a time dependent adjustment set. Robins (1987b) derives analogous sufficient conditions assuming the causal diagram represents a non-parametric structural equations model. See also Richardson and Robins (2013).

When $\mathbf{A}$ is a point intervention A, a time independent adjustment sets always exist. For instance, $\mathbf{Z} = \operatorname{pa}_{\mathcal{G}}(A)$ is one such set. However, for $\mathbf{A} = (A_0, \ldots, A_p)$ a joint intervention, a time independent covariate adjustment set $\mathbf{Z}$ may not exist in some graphs, as noted in Henckel et al. (2019). In contrast, a time dependent adjustment sets always exists, since $\mathbf{Z} \equiv (\mathbf{Z}_0, \mathbf{Z}_1, \ldots, \mathbf{Z}_p)$ where $\mathbf{Z}_0 \equiv \operatorname{pa}_{\mathcal{G}}(A_0)$ and $\mathbf{Z}_k \equiv \operatorname{pa}_{\mathcal{G}}(A_k) \setminus \left[\cup_{j=0}^{k-1} \operatorname{pa}_{\mathcal{G}}(A_j)\right], k = 1, \ldots, p$ is a time dependent adjustment set.

Example 1 *In the DAG of Figure 1, there is no time independent adjustment set relative to $(\mathbf{A}, Y)$ for $\mathbf{A} = (A_0, A_1)$. For instance, $\mathbf{Z} = (\mathbf{Z}_0, \mathbf{Z}_1)$ with $\mathbf{Z}_0 = \{L_0\}$ and $\mathbf{Z}_1 = \{L_1\}$, and $\widetilde{\mathbf{Z}} = \left(\widetilde{\mathbf{Z}}_0, \widetilde{\mathbf{Z}}_1\right)$, with $\widetilde{\mathbf{Z}}_0 = \{L_0\}$ and $\widetilde{\mathbf{Z}}_1 = \{L_1, U\}$, are two time dependent adjustment sets (Robins, 1987b).*

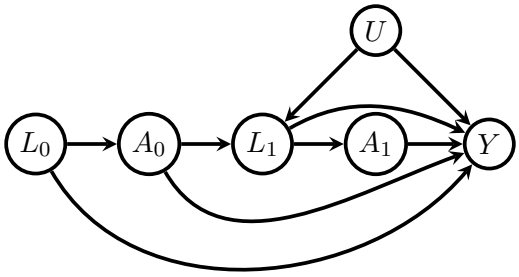

Figure 1: A DAG with two possible time dependent adjustment sets and no time independent adjustment sets.

We also have the following definition.

Definition 3 (Minimal covariate adjustment set) *Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$, let $\mathbf{A} \subset \mathbf{V}$ and $Y \in \mathbf{V} \backslash \mathbf{A}$. A set $\mathbf{Z} \subset \mathbf{V} \backslash \{\mathbf{A}, Y\}$ is a minimal time dependent (independent) adjustment set relative to $(\mathbf{A}, Y)$ in $\mathcal{G}$ if $\mathbf{Z}$ is a time dependent (independent) adjustment set and no proper subset of $\mathbf{Z}$ is a time dependent (independent) adjustment set.*

4. Non-parametric Estimation of an Interventional Mean

Suppose that $\mathbf{A}$ is a vector of variables taking values on a finite set $\mathcal{A}$ and one is interested in estimating some contrast

$$\Delta \equiv \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} E[Y_{\mathbf{a}}]$$

for given constants $c_{\mathbf{a}}$, $\mathbf{a} \in \mathcal{A}$. In particular if $\mathbf{A} = A$ is binary and $c_1 = 1$ and $c_0 = -1$ the preceding linear combination is equal to ATE. Suppose that, having postulated a causal graphical model, one finds that time independent adjustment sets exist. Having decided on one adjustment set $\mathbf{Z}$, one estimates

$$\Delta(P; \mathcal{G}) \equiv \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} \chi_{\mathbf{a}} (P; \mathcal{G}),$$

where

$$\chi_{\mathbf{a}} (P; \mathcal{G}) \equiv E_P [E_P [Y | \mathbf{A} = \mathbf{a}, \mathbf{Z}]] = E_P \left[\pi_{\mathbf{a}} (\mathbf{Z}; P)^{-1} I_{\mathbf{a}} (\mathbf{A}) Y \right],$$

by estimating each $\chi_{\mathbf{a}} (P; \mathcal{G})$ under a model $\mathcal{M}$ that makes at most smoothness or complexity assumptions on

$$b_{\mathbf{a}} (\mathbf{Z}; P) \equiv E_P [Y | \mathbf{A} = \mathbf{a}, \mathbf{Z}]$$

and/or

$$\pi_{\mathbf{a}} (\mathbf{Z}; P) \equiv P [\mathbf{A} = \mathbf{a} | \mathbf{Z}].$$

Examples of such estimating strategies include the inverse probability weighted estimator $\widehat{\chi}_{\mathbf{a}, IPW} = \mathbb{P}_n \left[\widehat{\pi}_{\mathbf{a}} (\mathbf{Z})^{-1} I_{\mathbf{a}} (\mathbf{A}) Y \right]$ where $\widehat{\pi}_{\mathbf{a}} (\cdot)$ is a series or kernel estimator of $P [\mathbf{A} = \mathbf{a} | \mathbf{Z} = \cdot]$

(Hirano et al., 2003), the outcome regression estimator $\mathbb{P}_n\left[\widehat{b}_{\mathbf{a}}(\mathbf{Z})\right]$ where $\widehat{b}_{\mathbf{a}}(\cdot)$ is a smooth estimator of $b_{\mathbf{a}}(\mathbf{Z};P)$ (Hahn, 1998) or the doubly-robust estimator (Van der Laan and Robins, 2003; Chernozhukov et al.; Smucler et al., 2019).

This estimation strategy effectively uses the causal model solely to provide guidance on the selection of the adjustment set but otherwise ignores the information about the interventional means $\chi_{\mathbf{a}}(P;\mathcal{G})$ encoded in the causal model. This is a strategy frequently followed in applications (Abadie and Cattaneo, 2018; Bottou et al., 2013; Hernan and Robins, 2019). It is well known (Robins et al., 1994) that estimators $\widehat{\chi}_{\mathbf{a},\mathbf{Z}}$ of $\chi_{\mathbf{a}}(P;\mathcal{G})$ based on the adjustment set $\mathbf{Z}$ that are regular and asymptotically linear under a model $\mathcal{M}$ that imposes at most smoothness or complexity assumptions on $b_{\mathbf{a}}(\mathbf{Z};P)$ and/or $\pi_{\mathbf{a}}(\mathbf{Z};P)$ have a unique influence function equal to

$$\psi_{P,\mathbf{a}}(\mathbf{Z};\mathcal{G}) \equiv \frac{I_{\mathbf{a}}(\mathbf{A})}{\pi_{\mathbf{a}}(\mathbf{Z};P)}(Y - b_{\mathbf{a}}(\mathbf{Z};P)) + b_{\mathbf{a}}(\mathbf{Z};P) - \chi_{\mathbf{a}}(P;\mathcal{G}), \tag{4}$$

where to avoid overloading the notation in $\psi_{P,\mathbf{a}}$ we do not explicitly write its dependence on $(Y,\mathbf{A})$.

Consequently, estimators $\widehat{\Delta}_{\mathbf{Z}} \equiv \sum_{\mathbf{a}\in\mathcal{A}} c_{\mathbf{a}}\widehat{\chi}_{\mathbf{a},\mathbf{Z}}$ of $\Delta(P;\mathcal{G})$ have a unique influence function equal to

$$\psi_{P,\Delta}(\mathbf{Z};\mathcal{G}) = \sum_{\mathbf{a}\in\mathcal{A}} c_{\mathbf{a}}\psi_{P,\mathbf{a}}(\mathbf{Z};\mathcal{G}).$$

For simplicity, we refer to asymptotically linear estimators of $\chi_{\mathbf{a}}(P;\mathcal{G})$ with influence function $\psi_{P,\mathbf{a}}(\mathbf{Z};\mathcal{G})$ as non-parametric estimators that use the adjustment set $\mathbf{Z}$ and we abbreviate them with NP-$\mathbf{Z}$.

The preceding discussion implies that any NP-$\mathbf{Z}$ estimator $\widehat{\chi}_{\mathbf{a},\mathbf{Z}}$ satisfies

$$\sqrt{n}\left\{\widehat{\chi}_{\mathbf{a},\mathbf{Z}} - \chi_{\mathbf{a}}(P;\mathcal{G})\right\} \xrightarrow{d} N\left(0, \sigma^2_{\mathbf{a},\mathbf{Z}}(P)\right)$$

where $\sigma^2_{\mathbf{a},\mathbf{Z}}(P) \equiv var_P\left[\psi_{P,\mathbf{a}}(\mathbf{Z};\mathcal{G})\right]$. Likewise, $\sqrt{n}\left\{\widehat{\Delta}_{\mathbf{Z}} - \Delta(P;\mathcal{G})\right\} \xrightarrow{d} N\left(0, \sigma^2_{\Delta,\mathbf{Z}}\right)$ where

$$\sigma^2_{\Delta,\mathbf{Z}}(P) \equiv var_P\left[\psi_{P,\Delta}(\mathbf{Z};\mathcal{G})\right].$$

Two natural questions of practical interest arise. The first is whether any two given time independent covariate adjustment sets, say $\mathbf{Z},\mathbf{Z}'$, are comparable in the sense that either

$$\sigma^2_{\Delta,\mathbf{Z}} \leq \sigma^2_{\Delta,\mathbf{Z}'} \text{ for all } P\in\mathcal{M}(\mathcal{G}) \text{ or } \sigma^2_{\Delta,\mathbf{Z}'} \leq \sigma^2_{\Delta,\mathbf{Z}} \text{ for all } P\in\mathcal{M}(\mathcal{G}).$$

The second is whether an optimal time independent adjustment set $\mathbf{O}$ exists such that for any other time independent adjustment set $\mathbf{Z}$,

$$\sigma^2_{\Delta,\mathbf{O}}(P) \leq \sigma^2_{\Delta,\mathbf{Z}}(P). \tag{5}$$

These questions were answered by Henckel et al. (2019) under (i) a linear causal graphical model, (ii) when $\Delta = E[Y_{\mathbf{a}} - Y_{\mathbf{a}'}]$ where $\mathbf{a}-\mathbf{a}'$ is the vector with all coordinates equal to zero except for coordinate j which is equal to one, and (iii) when Δ is estimated as the ordinary least squares estimator of the coefficient of A_j in the linear regression of Y on $\mathbf{A}$

and $\mathbf{Z}$ and $\sigma^2_{\Delta,\mathbf{Z}}$ is the asymptotic variance of such estimators. These authors showed that not all time independent covariate adjustment sets are comparable. However, they provided a graphical criterion to compare certain pairs of time independent covariate adjustment sets. They also provided a graphical criterion for characterizing the set $\mathbf{O}$, whenever a valid time independent covariate adjustment set exists. In particular, the criterion always returns an optimal valid time independent covariate adjustment set for $\mathbf{A} = A$ a point interventions.

In Section 5.1 we prove that the same graphical criteria remain valid for comparing time independent covariate adjustment sets and for characterizing the set $\mathbf{O}$ that satisfies (5) under an arbitrary, not necessarily linear, causal graphical model and for NP-$\mathbf{Z}$ estimators of an arbitrary contrast Δ. Moreover, for $\mathbf{A} = A$ a point intervention, we further show that there exists a minimal adjustment set $\mathbf{O}_{\min}$ included in $\mathbf{O}$ such that $\mathbf{O}_{\min}$ is optimal among the minimal adjustment sets; that is, for any other minimal adjustment set $\mathbf{Z}_{\min}$,

$$\sigma^2_{\Delta,\mathbf{O}_{\min}}(P) \leq \sigma^2_{\Delta,\mathbf{Z}_{\min}}(P), \tag{6}$$

where $\sigma^2_{\Delta,\mathbf{Z}_{\min}}(P)$ stands for either the asymptotic variance of the NP-$\mathbf{Z}_{min}$ estimator or the asymptotic variance of the OLS estimator of treatment effect of Henckel et al. (2019). In addition, we provide a graphical criterion for identifying $\mathbf{O}_{\min}$. Using the tools developed in van der Zander and Liskiewicz (2019), $\mathbf{O}$ and $\mathbf{O}_{\min}$ it can be shown that can be computed in polynomial time.

Consider next the case in which $\mathbf{A} = (A_0, \ldots, A_p)$ is a joint intervention with $p > 0$. In analogy with the time independent covariate adjustment case we consider in Section 5.2 the setting in which one uses the causal model to identify the collection of time dependent adjustment sets, but then for any given time dependent adjustment set $\mathbf{Z}$, one estimates each $E[Y_{\mathbf{a}}]$ ignoring the conditional independencies encoded in the causal graphical model. For instance, for $p = 1$, we study the asymptotic efficiency of estimators of

$$\begin{aligned}\chi_{a_0,a_1}(P;\mathcal{G}) &\equiv E_P\left\{E_P\left[E_P\left[Y|A_0 = a_0, A_1 = a_1, \mathbf{Z}_0, \mathbf{Z}_1\right]|A_0 = a_0, \mathbf{Z}_0\right]\right\} \\ &= E_P\left[\frac{I_{a_0}(A_0)}{P\left[A_0 = a_0|\mathbf{Z}_0\right]}\frac{I_{a_1}(A_1)}{P\left[A_1 = a_1|A_0 = a_0, \mathbf{Z}_0, \mathbf{Z}_1\right]}Y\right]\end{aligned}$$

for different time dependent adjustment sets $(\mathbf{Z}_0, \mathbf{Z}_1)$, under a model $\mathcal{M}$ that makes at most smoothness or complexity assumptions on

$$\begin{aligned}&b_{a_0,a_1}(\mathbf{Z}_0, \mathbf{Z}_1; P) \equiv E_P\left[Y|A_0 = a_0, A_1 = a_1, \mathbf{Z}_0, \mathbf{Z}_1\right], \\ &b_{a_0}(\mathbf{Z}_0; P) \equiv E_P\left[b_{a_0,a_1}(\mathbf{Z}_0, \mathbf{Z}_1; P)|A_0 = a, \mathbf{Z}_0\right]\end{aligned}$$

and/or

$$\begin{aligned}&\pi_{a_0,a_1}(\mathbf{Z}_0, \mathbf{Z}_1; P) \equiv P\left[A_1 = a_1|A_0 = a_0, \mathbf{Z}_0, \mathbf{Z}_1\right], \\ &\pi_{a_0}(\mathbf{Z}_0; P) \equiv P\left[A_0 = a_0|\mathbf{Z}_0\right].\end{aligned}$$

See Van der Laan and Robins (2003). Just as for the case of time independent adjustment sets, not all time dependent adjustment sets are comparable in terms of their asymptotic variance uniformly for all $P \in \mathcal{M}(\mathcal{G})$. However, in Section 5.2 we generalize the aforementioned graphical criterion that allows the comparison of certain time dependent adjustment sets. Nevertheless we show by example that unlike the case of time independent adjustment sets, even though a time dependent adjustment set always exists, there are DAGs in which no uniformly optimal time dependent adjustment set exists.

5. Comparison of Adjustment Sets

In Section 5.1 we show that the graphical criteria for comparing time independent adjustment sets and for identifying the optimal time independent adjustment set of Henckel et al. (2019) is valid also when treatment effects are estimated non-parametrically. In Section 5.2 we provide results for time dependent adjustment sets.

5.1 Time Independent Adjustment Sets

Lemma 4 (Supplementation with time independent precision variables) *Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$, let $\mathbf{A} \subset \mathbf{V}$ and $Y \in \mathbf{V} \setminus \mathbf{A}$ with $\mathbf{A}$ a random vector taking values on a finite set. Suppose $\mathbf{B} \subset \mathbf{V} \setminus \{\mathbf{A}, Y\}$ is a time independent adjustment set relative to $(\mathbf{A}, Y)$ in $\mathcal{G}$ and suppose $\mathbf{G}$ is a disjoint set with $\mathbf{B}$ that satisfies*

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{G} \mid \mathbf{B}.$$

Then $(\mathbf{G}, \mathbf{B})$ is also a time independent adjustment set relative to $(\mathbf{A}, Y)$ in $\mathcal{G}$ and for all $P \in \mathcal{M}(\mathcal{G})$

$$\sigma^2_{\mathbf{a},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{G},\mathbf{B}}(P) = E_P\left[\left\{\frac{1}{\pi_{\mathbf{a}}(\mathbf{B};P)} - 1\right\} var_P\left[b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)|\,\mathbf{B}\right]\right] \geq 0. \tag{7}$$

Furthermore,

$$\sigma^2_{\Delta,\mathbf{B}}(P) - \sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P) = \mathbf{c}^T var_P(\mathbf{Q})\,\mathbf{c} \geq 0$$

where $\mathbf{c} \equiv (c_{\mathbf{a}})_{\mathbf{a}\in\mathbf{A}}$ and $\mathbf{Q} \equiv [Q_{\mathbf{a}}]_{\mathbf{a}\in\mathbf{A}}$ with

$$\begin{aligned}
&Q_{\mathbf{a}} \equiv \left\{\frac{I_{\mathbf{a}}(\mathbf{A})}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)} - 1\right\}\{b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P) - b_{\mathbf{a}}(\mathbf{B};P)\},\\
&var_P(Q_{\mathbf{a}}) = E_P\left[\left\{\frac{1}{\pi_{\mathbf{a}}(\mathbf{B};P)} - 1\right\} var_P(b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P) \mid \mathbf{B})\right],\\
&\text{and } cov_P[Q_{\mathbf{a}}, Q_{\mathbf{a}'}] = -E_P[cov_P\{b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P), b_{\mathbf{a}'}(\mathbf{G},\mathbf{B};P)|\mathbf{B}\}] \text{ for } \mathbf{a} \neq \mathbf{a}'.
\end{aligned}$$

In particular,

$$\begin{aligned}
\sigma^2_{ATE,\mathbf{B}}(P) - \sigma^2_{ATE,\mathbf{G},\mathbf{B}}(P) \;=\;& E_P\left[\left\{\frac{1}{\pi_{a=1}(\mathbf{B};P)} - 1\right\} var_P(b_{a=1}(\mathbf{G},\mathbf{B};P) \mid \mathbf{B})\right]\\
&+E_P\left[\left\{\frac{1}{\pi_{a=0}(\mathbf{B};P)} - 1\right\} var_P(b_{a=0}(\mathbf{G},\mathbf{B};P) \mid \mathbf{B})\right]\\
&-2E_P[cov_P\{b_{a=1}(\mathbf{G},\mathbf{B};P), b_{a=0}(\mathbf{G},\mathbf{B};P)|\mathbf{B}\}]\\
&\geq 0.
\end{aligned}$$

For the special case in which $\mathbf{B} = \emptyset$, formula (7) was derived in Robins and Rotnitzky (1992) and Hahn (1998). The formula quantifies the reduction in variance associated with supplementing an adjustment set with 'precision' variables, i.e., variables that may help predict the outcome within treatment levels but are not associated with treatments after controlling for the already existing adjustment set. Notice that $var_P[b_a(\mathbf{G},\mathbf{B};P)|\,\mathbf{B}]$ quantifies the additional explanatory power carried by $\mathbf{G}$ for Y after adjusting for $\mathbf{B}$. In the

DAG represented in Figure 2, $\mathbf{B} = \{B\}$ and $\mathbf{G} = \{G\}$ satisfy the conditions of Lemma 4. In that DAG, $var_P\left[b_a(\mathbf{G}, \mathbf{B}; P) | \mathbf{B}\right]$ increases as the strength of the association encoded in the red edge increases and the one encoded in the green edge decreases. In contrast, $\{1/\pi_a(\mathbf{B}; P) - 1\}$ is always greater than 0, and it is more variable, and thus tends to have larger values, the stronger the marginal association of $\mathbf{B}$ with A. In the DAG in Figure 2, this association is represented by the blue edge.

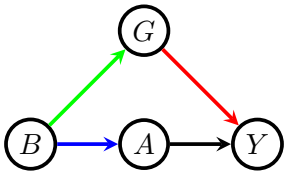

Figure 2: A DAG illustrating Lemmas 4 and 5.

Lemma 5 (Deletion of time independent overadjustment variables) *Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$, let $\mathbf{A} \subset \mathbf{V}$ and $Y \in \mathbf{V} \setminus \mathbf{A}$ with $\mathbf{A}$ a random vector taking values on a finite set. Suppose $(\mathbf{G} \cup \mathbf{B}) \subset \mathbf{V} \setminus \{\mathbf{A}, Y\}$ is a time independent adjustment set relative to $(\mathbf{A}, Y)$ in $\mathcal{G}$ with $\mathbf{G}$ and $\mathbf{B}$ disjoint and suppose*

$$Y \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{G}, \mathbf{A}.$$

Then $\mathbf{G}$ is also an adjustment set relative to $(\mathbf{A}, Y)$ in $\mathcal{G}$ and for all $P \in \mathcal{M}(\mathcal{G})$

$$\begin{aligned}
&\sigma^2_{\mathbf{a},\mathbf{G},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{G}}(P) = \\
&E_P\left[\pi_{\mathbf{a}}(\mathbf{G}; P)\, var_P(Y | \mathbf{A} = \mathbf{a}, \mathbf{G})\, var_P\left(\frac{1}{\pi_{\mathbf{a}}(\mathbf{G}, \mathbf{B}; P)}\middle| \mathbf{A} = \mathbf{a}, \mathbf{G}\right)\right] \geq 0. \qquad (8)
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P) - \sigma^2_{\Delta,\mathbf{B}}(P) = \\
&\sum_{a \in \mathcal{A}} c^2_{\mathbf{a}} E_P\left\{\pi_{\mathbf{a}}(\mathbf{G}; P) var_P(Y \mid \mathbf{A} = \mathbf{a}, \mathbf{G}) var_P\left[\frac{1}{\pi_{\mathbf{a}}(\mathbf{G}, \mathbf{B}; P)} \mid \mathbf{A} = \mathbf{a}, \mathbf{G}\right]\right\} \geq 0.
\end{aligned}$$

In particular,

$$\begin{aligned}
&\sigma^2_{ATE,\mathbf{B}}(P) - \sigma^2_{ATE,\mathbf{G},\mathbf{B}}(P) = \\
&E_P\left\{\pi_{a=0}(\mathbf{G}; P) var_P(Y \mid A = 0, \mathbf{G}) var_P\left[\frac{1}{\pi_{a=0}(\mathbf{G}, \mathbf{B}; P)} \mid A = 0, \mathbf{G}\right]\right\} + \\
&E_P\left\{\pi_{a=1}(\mathbf{G}; P) var_P(Y \mid A = 1, \mathbf{G}) var_P\left[\frac{1}{\pi_{a=0}(\mathbf{G}, \mathbf{B}; P)} \mid A = 1, \mathbf{G}\right]\right\} \geq 0.
\end{aligned}$$

Formula (8) quantifies the increase in variance incurred by keeping 'overadjustment' variables that are marginally associated with treatment but that do not help predict the outcome within levels of treatment and the remaining adjusting variables. This result

extends to the non-parametric setting the well understood increase in variance induced by adding covariates that have no partial correlation with the outcome in a linear regression model (Cochran, 1968). This feature has also been noticed in a number of non-linear regression settings (Mantel and Haenszel, 1959; Breslow, 1982; Gail, 1988; Robinson and Jewell, 1991; Neuhaeuser and Becher, 1997; De Stavola and Cox, 2008).

Notice that $var_P(Y|A=a,\mathbf{G})$ is zero if $\mathbf{G}$ is a perfect predictor of Y. In such extreme case, the formula indicates that it is irrelevant whether one keeps the overadjustment variables $\mathbf{B}$. In general, $\mathbf{B}$ is more harmful the weaker the association between $\mathbf{G}$ and Y within levels of A is. For example, in the causal diagram in Figure 2, the penalty for keeping overadjustment variables increases as the strength of the association represented in the red arrow decreases. Furthermore, the quantity $var_P\left(1/\pi_a\left(\mathbf{G},\mathbf{B};P\right)|\, A=a,\mathbf{G}\right)$ indicates that $\mathbf{B}$ is also more harmful the weaker the association between $\mathbf{G}$ and $\mathbf{B}$ within levels of A, and the stronger the association between $\mathbf{B}$ and A within levels of $\mathbf{G}$. For instance, in the causal diagram in Figure 2, $\mathbf{B}$ is also more harmful the weaker the association represented by the green arrow is and the stronger the association represented by the blue arrow is.

Theorem 6 *Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$, let $\mathbf{A}\subset\mathbf{V}$ and $Y\in\mathbf{V}\setminus\mathbf{A}$ with $\mathbf{A}$ a random vector taking values on a finite set. Suppose $\mathbf{G}\subset\mathbf{V}\setminus\{\mathbf{A},\mathbf{Y}\}$ and $\mathbf{B}\subset\mathbf{V}\setminus\{\mathbf{A},Y\}$ are two time independent adjustment sets relative to $(\mathbf{A},Y)$ in $\mathcal{G}$ such that*

$$\mathbf{A}\perp\!\!\!\perp_{\mathcal{G}}\ [\mathbf{G}\backslash\mathbf{B}]\ |\ \mathbf{B} \tag{9}$$

$$Y\perp\!\!\!\perp_{\mathcal{G}}\ [\mathbf{B}\backslash\mathbf{G}]\ |\ \mathbf{G},\mathbf{A}. \tag{10}$$

Then $\sigma^2_{\mathbf{a},\mathbf{B}}(P)-\sigma^2_{\mathbf{a},\mathbf{G}}(P)\geq 0$ for all $P\in\mathcal{M}(\mathcal{G})$. Specifically,

$$\begin{aligned}&\sigma^2_{\mathbf{a},\mathbf{B}}(P)-\sigma^2_{\mathbf{a},\mathbf{G}}(P)=\\&E_P\left[\left\{\frac{1}{\pi_{\mathbf{a}}(\mathbf{B};P)}-1\right\}var_P\left[b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)|\,\mathbf{B}\right]\right]+\\&E_P\left[\pi_{\mathbf{a}}(\mathbf{G};P)\,var_P(Y|\mathbf{A}=\mathbf{a},\mathbf{G})\,var_P\left(\left.\frac{1}{\pi_a(\mathbf{G},\mathbf{B};P)}\right|\mathbf{A}=\mathbf{a},\mathbf{G}\right)\right].\end{aligned}$$

Moreover, $\sigma^2_{\Delta,\mathbf{B}}(P)-\sigma^2_{\Delta,\mathbf{G}}(P)\geq 0$ for all $P\in\mathcal{M}(\mathcal{G})$. Specifically,

$$\begin{aligned}&\sigma^2_{\Delta,\mathbf{B}}(P)-\sigma^2_{\Delta,\mathbf{G}}(P)=\\&\mathbf{c}^T var_P(\mathbf{Q})\,\mathbf{c}+\\&\sum_{a\in\mathcal{A}}c^2_{\mathbf{a}}E_P\left\{\pi_{\mathbf{a}}(\mathbf{G};P)var_P(Y\mid\mathbf{A}=\mathbf{a},\mathbf{G})var_P\left[\frac{1}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)}\mid\mathbf{A}=\mathbf{a},\mathbf{G}\right]\right\},\end{aligned}$$

where $\mathbf{Q}$ *is defined as in Lemma 4. In particular,* $\sigma^2_{ATE,\mathbf{B}}(P) - \sigma^2_{ATE,\mathbf{G}}(P) \geq 0$ *for all* $P \in \mathcal{M}(\mathcal{G})$*. Specifically,*

$$
\begin{aligned}
&\sigma^2_{ATE,\mathbf{B}}(P) - \sigma^2_{ATE,\mathbf{G}}(P) = \\
&E_P\left[\left\{\frac{1}{\pi_{a=1}(\mathbf{B};P)} - 1\right\} var_P(b_{a=1}(\mathbf{G},\mathbf{B};P) \mid \mathbf{B})\right] + \\
&E_P\left[\left\{\frac{1}{\pi_{a=0}(\mathbf{B};P)} - 1\right\} var_P(b_{a=0}(\mathbf{G},\mathbf{B};P) \mid \mathbf{B})\right] - \\
&2E_P\left[cov_P\left\{b_{a=1}(\mathbf{G},\mathbf{B};P), b_{a=0}(\mathbf{G},\mathbf{B};P)|\mathbf{B}\right\}\right] + \\
&E_P\left\{\pi_{a=0}(\mathbf{G};P) var_P(Y \mid A=0,\mathbf{G}) var_P\left[\frac{1}{\pi_{a=0}(\mathbf{G},\mathbf{B};P)} \mid A=0,\mathbf{G}\right]\right\} + \\
&E_P\left\{\pi_{a=1}(\mathbf{G};P) var_P(Y \mid A=1,\mathbf{G}) var_P\left[\frac{1}{\pi_{a=0}(\mathbf{G},\mathbf{B};P)} \mid A=1,\mathbf{G}\right]\right\}.
\end{aligned}
$$

Proof Write $\sigma^2_{\mathbf{a},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{G}}(P) = \sigma^2_{\mathbf{a},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{B}\cup(\mathbf{G}\setminus\mathbf{B})}(P) + \sigma^2_{\mathbf{a},\mathbf{G}\cup(\mathbf{B}\setminus\mathbf{G})}(P) - \sigma^2_{\mathbf{a},\mathbf{G}}(P)$ and apply Lemmas 4 and 5. The derivations for the expressions for $\sigma^2_{\Delta,\mathbf{B}}(P) - \sigma^2_{\Delta,\mathbf{G}}(P)$ and $\sigma^2_{ATE,\mathbf{B}}(P) - \sigma^2_{ATE,\mathbf{G}}(P)$ are similar. ■

The preceding theorem provides an intuitive decomposition for the gain in efficiency of using adjustment set $\mathbf{G}$ as opposed to set $\mathbf{B}$. The difference $\sigma^2_{\mathbf{a},\mathbf{B}} - \sigma^2_{\mathbf{a},\mathbf{B}\cup(\mathbf{G}\setminus\mathbf{B})}$ represents the gain due to supplementing $\mathbf{B}$ with the precision component $\mathbf{G}\setminus\mathbf{B}$ and $\sigma^2_{\mathbf{a},\mathbf{G}\cup(\mathbf{B}\setminus\mathbf{G})} - \sigma^2_{\mathbf{a},\mathbf{G}}$ represents the gain from removing from $\mathbf{G}\cup\mathbf{B}$ the overadjustment component $\mathbf{B}\setminus\mathbf{G}$.

Theorem 6 is analogous to Theorem 3.10 from Henckel et al. (2019), except that it is valid for arbitrary causal graphical models, instead of causal linear models, and for NP-$\mathbf{Z}$ estimators of treatment effects instead of ordinary least squares estimators. Likewise, Lemmas 4 and 5 are analogous to Henckel et al's Corollaries 3.4 and 3.5. Building on their Corollary 3.5, Henckel et al. (2019) provided a simple procedure that, for a valid adjustment set, returns a pruned valid adjustment set that yields OLS estimators of treatment effects with smaller asymptotic variance. Because the validity of their pruning procedure relies only on the ordering of the asymptotic variances corresponding to two adjustment sets implied by the d-separation assumptions of their Corollary 3.5, and because the same ordering of the adjustment sets is valid for the variances of the corresponding NP-$\mathbf{Z}$ estimators, then we conclude that the pruning algorithm of Henckel et al. (2019) also returns a pruned valid adjustment set that yields an NP-$\mathbf{Z}$ estimator of treatment effect with smaller asymptotic variance.

As noted by Henckel et al. (2019), not all pairs of valid time independent adjustment sets can be ordered using the d-separation conditions in Theorem 6. In fact, there exist DAGs $\mathcal{G}$ with time independent adjustment sets $\mathbf{Z}$ and $\widetilde{\mathbf{Z}}$ for which $\sigma^2_{a,\mathbf{Z}}(P) > \sigma^2_{a,\widetilde{\mathbf{Z}}}(P)$ for some $P \in \mathcal{M}(\mathcal{G})$ and $\sigma^2_{a,\widetilde{\mathbf{Z}}}(P') > \sigma^2_{a,\mathbf{Z}}(P')$ for some other $P' \in \mathcal{M}(\mathcal{G})$ as the following example illustrates.

Example 2 *In the DAG in Figure 3,* $\mathbf{Z} = \{O_1, W_2\}$ *and* $\widetilde{\mathbf{Z}} = \{O_2, W_1\}$ *are time independent adjustment sets relative to* (A, Y)*. The adjustment set* $\mathbf{Z}$ *yields a smaller asymptotic variance than the adjustment set* $\widetilde{\mathbf{Z}}$ *if the association encoded in the green edge is stronger than*

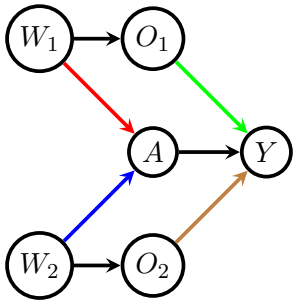

Figure 3: A DAG with two time independent adjustment sets, $\mathbf{Z}=\{O_1, W_2\}$ and $\widetilde{\mathbf{Z}}=\{O_2, W_1\}$, that cannot be compared. Note that $\mathbf{Z}$ and $\widetilde{\mathbf{Z}}$ are minimal time independent adjustment sets.

that in the brown edge and the one encoded in the blue edge is weaker than the one in the red edge. By symmetry, the adjustment set $\widetilde{\mathbf{Z}}$ is more efficient than $\mathbf{Z}$ if the words stronger and weaker are interchanged in the preceding sentence. Henckel et al. (2019) illustrated the impossibility of ordering all time independent adjustment sets by the asymptotic variances of the corresponding adjusted linear estimators with a diagram different from the one in Figure 3, in which the treatment was unconfounded.

Following Henckel et al. (2019) we let $\mathrm{cn}(\mathbf{A}, Y, \mathcal{G})$ be the set of all nodes that lie on a causal path between a node in $\mathbf{A}$ and Y and are not equal to any node in $\mathbf{A}$ and we define the forbidden set as

$$\mathrm{forb}(\mathbf{A}, Y, \mathcal{G}) \equiv \mathrm{de}_{\mathcal{G}}\left(\mathrm{cn}(\mathbf{A}, Y, \mathcal{G})\right) \cup \{\mathbf{A}\}.$$

Also,

$$\mathbf{O}(\mathbf{A}, Y, \mathcal{G}) \equiv \mathrm{pa}_{\mathcal{G}}\left(\mathrm{cn}(\mathbf{A}, Y, \mathcal{G})\right) \setminus \mathrm{forb}(\mathbf{A}, Y, \mathcal{G}).$$

Henckel et al. (2019) showed that, if a time independent adjustment set relative to $(\mathbf{A}, Y)$ in $\mathcal{G}$ exists, then $\mathbf{O}(\mathbf{A}, Y, \mathcal{G})$ satisfies the graphical necessary and sufficient conditions of Shpitser et al. (2010) to be an adjustment set. Furthermore, Lemmas E.4 and E.5 of Henckel et al. (2019) showed that the conditions (9) and (10) hold for $\mathbf{G} = \mathbf{O}(\mathbf{A}, Y, \mathcal{G})$ and $\mathbf{B}$ any adjustment set. Consequently, we have the following important corollary to Theorem 6.

Theorem 7 *Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$, let $\mathbf{A} \subset \mathbf{V}$ and $Y \in \mathbf{V} \setminus \mathbf{A}$ with $\mathbf{A}$ a random vector taking values on a finite set. If a valid time independent adjustment set $\mathbf{Z}$ relative to $(\mathbf{A}, Y)$ in $\mathcal{G}$ exists then $\mathbf{O} = \mathbf{O}(\mathbf{A}, Y, \mathcal{G})$ is a time independent adjustment set and $\sigma^2_{\mathbf{a},\mathbf{Z}}(P) - \sigma^2_{\mathbf{a},\mathbf{O}}(P) \geq 0$ for all $P \in \mathcal{M}(\mathcal{G})$. Specifically,*

$$\begin{aligned}
&\sigma^2_{\mathbf{a},\mathbf{Z}}(P) - \sigma^2_{\mathbf{a},\mathbf{O}}(P) = \\
&E_P\left[\left\{\frac{1}{\pi_{\mathbf{a}}(\mathbf{Z}; P)} - 1\right\} var_P\left[b_{\mathbf{a}}(\mathbf{O}, \mathbf{Z}; P) \middle| \mathbf{Z}\right]\right] + \\
&E_P\left[\pi_{\mathbf{a}}(\mathbf{O}; P)\, var_P(Y | A = a, \mathbf{O})\, var_P\left(\frac{1}{\pi_{\mathbf{a}}(\mathbf{O}, \mathbf{Z}; P)} \middle| \mathbf{A} = \mathbf{a}, \mathbf{O}\right)\right]
\end{aligned}$$

and the corresponding formulae for Δ and ATE hold.

Corollary 8 *If $\mathbf{A} = A$ is point intervention then $\mathbf{O}(A, Y, \mathcal{G})$ is an optimal valid time independent adjustment set.*

Corollary 8 follows immediately from Theorem 7 and the fact that $\text{pa}_{\mathcal{G}}(A)$ is always a valid time independent adjustment set relative to (A, Y) in $\mathcal{G}$.

Example 3 *In the DAG of Figure 3, $\mathbf{O}(A, Y, \mathcal{G}) = (O_1, O_2)$ is the optimal adjustment set.*

van der Zander and Liskiewicz (2019) proposed an algorithm that, given a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, computes $\mathbf{O}(\mathbf{A}, Y, \mathcal{G})$ with worst-case complexity $\mathcal{O}(|\mathbf{V}| + |\mathbf{E}|)$, where $|\mathbf{V}|$ is the number of nodes in $\mathcal{G}$ and $|\mathbf{E}|$ is the number of edges in $\mathcal{G}$.

For simplicity, from now on when no confusion can arise, we abbreviate $\mathbf{O} \equiv \mathbf{O}(\mathbf{A}, Y, \mathcal{G})$.

An interesting question is whether one can find an optimal adjustment set among the minimal adjustment sets. In the next theorem we show that such adjustment exists for point interventions. Specifically, let $\mathbf{A} = A$ be a point intervention and let $\mathbf{O}_{\min} \subset \mathbf{O}$ be the subset of $\mathbf{O}$ with the smallest number of vertices such that

$$A \perp\!\!\!\perp_{\mathcal{G}} [\mathbf{O} \backslash \mathbf{O}_{\min}] \,|\mathbf{O}_{\min}.$$

The graphoid properties of d-separation (Lauritzen, 1996) imply that $\mathbf{O}_{\min}$ is unique. For completeness we provide a proof of this result in Lemma 22 in the Appendix. Note that $\mathbf{O}_{\min}$ is empty when the empty set is a valid time independent adjustment set. The next theorem establishes that $\mathbf{O}_{\min}$ is a minimal adjustment set relative to (A, Y) in $\mathcal{G}$. Furthermore, it establishes that it is optimal among all minimal adjustment sets.

Theorem 9 *Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$, let A and Y be two distinct vertices in $\mathbf{V}$ with A corresponding to a point intervention.*

1. *$\mathbf{O}_{\min}$ as defined above is a minimal adjustment set relative to (A, Y) in $\mathcal{G}$.*
2. *If $\mathbf{Z}_{\min}$ is another minimal adjustment set relative to (A, Y) in $\mathcal{G}$ then,*

$$A \perp\!\!\!\perp_{\mathcal{G}} [\mathbf{O}_{\min} \backslash \mathbf{Z}_{\min}] \mid \mathbf{Z}_{\min} \quad \textit{and} \quad Y \perp\!\!\!\perp_{\mathcal{G}} [\mathbf{Z}_{\min} \backslash \mathbf{O}_{\min}] \mid \mathbf{O}_{\min}, A.$$

   Consequently, for all $P \in \mathcal{M}(\mathcal{G})$, $\sigma^2_{a,\mathbf{Z}_{\min}}(P) - \sigma^2_{a,\mathbf{O}_{\min}}(P) \geq 0$. Specifically,

$$\sigma^2_{a,\mathbf{Z}_{\min}}(P) - \sigma^2_{a,\mathbf{O}_{\min}}(P) =$$
$$E_P\left[\left\{\frac{1}{\pi_a(\mathbf{Z}_{\min}; P)} - 1\right\} var_P\left[b_a(\mathbf{O}_{\min}, \mathbf{Z}_{\min}; P) \middle| \mathbf{Z}_{\min}\right]\right] +$$
$$E_P\left[\pi_a(\mathbf{O}_{\min}; P)\, var_P(Y | A = a, \mathbf{O}_{\min})\, var_P\left(\frac{1}{\pi_a(\mathbf{O}_{\min}, \mathbf{Z}_{\min}; P)} \middle| A = a, \mathbf{O}_{\min}\right)\right]$$

   and the corresponding formulae hold for Δ and ATE.

3. *For any minimal adjustment set $\mathbf{Z}_{\min}$, $\mathbf{Z}_{\min} \cap [\mathbf{O} \backslash \mathbf{O}_{\min}] = \emptyset$.*

Remark 10 *Note that the conclusions one and two of Theorem 9 are purely graphical. Therefore, invoking Theorem 3.1 of Henckel et al. (2019), we conclude that $\mathbf{O}_{min}$ is also the minimal adjustment set that yields the adjusted OLS estimators of treatment effects with smallest variance among all adjusted OLS estimators of treatment effects that adjust for minimal adjustment sets.*

5.2 Time Dependent Adjustment Sets

Suppose that $\mathbf{A} = (A_0, \ldots, A_p)$ is a joint intervention and for a given time dependent adjustment set $\mathbf{Z} = (\mathbf{Z}_0, \ldots, \mathbf{Z}_p)$ in order to estimate a given contrast $\Delta(P; \mathcal{G}) \equiv \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} E[Y_{\mathbf{a}}]$ one estimates each interventional mean

$$E\left[Y_{\mathbf{a}}\right] = E_P\left\{\left\{E_P\left\{E_P\left[E_P\left[Y|\mathbf{a}, \mathbf{Z}\right]|\overline{\mathbf{a}}_{p-1}, \overline{\mathbf{Z}}_{p-1}\right]|\overline{\mathbf{a}}_{p-2}, \overline{\mathbf{Z}}_{p-2}\right\} \cdots | \, \mathbf{a}_0, \mathbf{Z}_0\right\}\right\} \equiv \chi_{\mathbf{a}}(P; \mathcal{G}),$$

ignoring the conditional independencies encoded in the causal graphical model, and making at most smoothness or complexity assumptions on the iterated conditional means

$$b_{\overline{\mathbf{a}}_j}(\overline{\mathbf{Z}}_j; P) \equiv E_P\left\{E_P\left\{E_P\left[E_P\left[Y|\mathbf{a}, \mathbf{Z}\right]|\overline{\mathbf{a}}_{p-1}, \overline{\mathbf{Z}}_{p-1}\right]|\overline{\mathbf{a}}_{p-2}, \overline{\mathbf{Z}}_{p-2}\right\} \cdots | \, \overline{\mathbf{a}}_j, \overline{\mathbf{Z}}_j\right\}$$

and/or on the conditional treatment probabilities

$$\pi_{a_j}(\overline{\mathbf{Z}}_j; P) \equiv P\left(A_j = a_j \mid \overline{\mathbf{A}}_{j-1} = \overline{\mathbf{a}}_{j-1}, \overline{\mathbf{Z}}_j\right).$$

It is well known (Robins and Rotnitzky, 1995) that estimators $\widehat{\chi}_{\mathbf{a},\mathbf{Z}}$ of $\chi_{\mathbf{a}}(P; \mathcal{G})$ that are regular and asymptotically linear under a model $\mathcal{M}$ that imposes at most smoothness or complexity assumptions on $b_{\overline{\mathbf{a}}_j}$ and/or π_{a_j} have a unique influence function equal to

$$\psi_{P,\mathbf{a}}(\mathbf{Z}; P) \equiv \frac{I_{\mathbf{a}}(\mathbf{A})}{\lambda_{\overline{\mathbf{a}}_p}(\mathbf{Z}; P)}\{Y - \chi_{\mathbf{a}}(P; \mathcal{G})\} - \sum_{k=0}^{p} g_k(\overline{\mathbf{A}}_k, \overline{\mathbf{Z}}_k; P), \tag{11}$$

where

$$g_k(\overline{\mathbf{A}}_k, \overline{\mathbf{Z}}_k; P) = \frac{I_{\overline{\mathbf{a}}_{k-1}}(\overline{\mathbf{A}}_{k-1})}{\lambda_{\overline{\mathbf{a}}_{k-1}}(\overline{\mathbf{Z}}_{k-1}; P)}\left\{\frac{I_{a_k}(A_k)}{\pi_{a_k}(\overline{\mathbf{Z}}_k; P)} - 1\right\}\left\{b_{\overline{\mathbf{a}}_k}(\overline{\mathbf{Z}}_k; P) - \chi_{\mathbf{a}}(P; \mathcal{G})\right\}$$

with $\lambda_{\overline{\mathbf{a}}_{k-1}}(\overline{\mathbf{Z}}_{k-1}; P) \equiv \prod_{j=0}^{k-1} \pi_{a_j}(\overline{\mathbf{Z}}_j; P)$ and $I_{\overline{\mathbf{a}}_{-1}}(\overline{\mathbf{A}}_{-1})\left[\lambda_{\overline{\mathbf{a}}_{-1}}(\overline{\mathbf{Z}}_{-1}; P)\right]^{-1} \equiv 1$.

Hence, regular and asymptotically linear estimators $\widehat{\Delta}_{\mathbf{Z}} \equiv \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} \widehat{\chi}_{\mathbf{a},\mathbf{Z}}$ of $\Delta(P; \mathcal{G})$ have a unique influence function equal to $\psi_{P,\Delta}(\mathbf{Z}; \mathcal{G}) = \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} \psi_{P,\mathbf{a}}(\mathbf{Z}; \mathcal{G})$. Therefore $\sqrt{n}\left\{\widehat{\chi}_{\mathbf{a},\mathbf{Z}} - \chi_{\mathbf{a}}(P; \mathcal{G})\right\} \xrightarrow{d} N\left(0, \sigma^2_{\mathbf{a},\mathbf{Z}}(P)\right)$ where $\sigma^2_{\mathbf{a},\mathbf{Z}}(P) \equiv var_P\left[\psi_{P,\mathbf{a}}(\mathbf{Z}; \mathcal{G})\right]$. Likewise, $\sqrt{n}\left\{\widehat{\Delta}_{\mathbf{Z}} - \Delta(P; \mathcal{G})\right\} \xrightarrow{d} N\left(0, \sigma^2_{\Delta,\mathbf{Z}}\right)$ where $\sigma^2_{\Delta,\mathbf{Z}}(P) \equiv var_P\left[\psi_{P,\Delta}(\mathbf{Z}; \mathcal{G})\right]$.

The following lemmas extend Lemmas 4 and 5 from time independent adjustment sets to time dependent adjustment sets. Throughout we let $b_{\overline{\mathbf{a}}_{-1}}\left(\overline{\mathbf{G}}_{-1}, \overline{\mathbf{B}}_{-1}; P\right) \equiv \chi_{\mathbf{a}}(P; \mathcal{G})$.

Lemma 11 (Supplementation with time dependent precision variables) *Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$, let $\mathbf{A} = (A_0, \ldots, A_p)$ be a topologically ordered vertex set in $\mathbf{V}$ disjoint with $Y \in \mathbf{V}$. Assume $A_j, j = 0, \ldots, p$, correspond to finite valued random variables. Suppose*

$$\mathbf{B} = (\mathbf{B}_0, \ldots, \mathbf{B}_p) \subset \mathbf{V} \backslash \{\mathbf{A}, Y\}$$

is a time dependent adjustment set relative to $(\mathbf{A}, Y)$ in $\mathcal{G}$ and suppose $\mathbf{G} = (\mathbf{G}_0, \ldots, \mathbf{G}_p)$ is a set disjoint with $\mathbf{B}$ that satisfies

$$A_j \perp\!\!\!\perp_{\mathcal{G}} \overline{\mathbf{G}}_j \mid \overline{\mathbf{B}}_j, \overline{\mathbf{A}}_{j-1} \text{ for } j = 0, \ldots, p, \tag{12}$$

where $\overline{\mathbf{A}}_{-1} = \emptyset$. *Then* $(\mathbf{G}, \mathbf{B}) = [(\mathbf{G}_0, \mathbf{B}_0), (\mathbf{G}_1, \mathbf{B}_1), \ldots, (\mathbf{G}_p, \mathbf{B}_p)]$ *is also an time dependent adjustment set relative to* $(\mathbf{A}, Y)$ *in* $\mathcal{G}$ *and for all* $P \in \mathcal{M}(\mathcal{G})$,

$$\sigma^2_{\mathbf{a},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{G},\mathbf{B}}(P) \geq 0 \quad \textit{and} \quad \sigma^2_{\Delta,\mathbf{B}}(P) - \sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P) \geq 0.$$

In the Appendix we provide formulas for $\sigma^2_{\mathbf{a},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{G},\mathbf{B}}(P)$ *and* $\sigma^2_{\Delta,\mathbf{B}}(P) - \sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P)$.

Lemma 12 (Deletion of time dependent overadjustment variables)
Let $\mathcal{G}$ *be a DAG with vertex set* $\mathbf{V}$, *let* $\mathbf{A} = (A_0, \ldots, A_p)$ *be a topologically ordered vertex set in* $\mathbf{V}$ *disjoint with* $Y \in \mathbf{V}$. *Assume* $A_j, j = 0, \ldots, p$, *correspond to finite valued random variables. Suppose* $(\mathbf{G}, \mathbf{B}) \equiv [(\mathbf{G}_0, \mathbf{B}_0), (\mathbf{G}_1, \mathbf{B}_1), \ldots, (\mathbf{G}_p, \mathbf{B}_p)] \subset \mathbf{V} \backslash \{\mathbf{A}, Y\}$ *is a time dependent adjustment set relative to* $(\mathbf{A}, Y)$ *in* $\mathcal{G}$ *with* $\mathbf{G}$ *and* $\mathbf{B}$ *disjoint and suppose that*

$$Y \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{G}, \mathbf{A}. \tag{13}$$

and

$$\mathbf{G}_j \perp\!\!\!\perp_{\mathcal{G}} \overline{\mathbf{B}}_{j-1} \mid \overline{\mathbf{G}}_{j-1}, \overline{\mathbf{A}}_{j-1} \textit{ for } j = 1, \ldots, p. \tag{14}$$

Then $\mathbf{G} = (\mathbf{G}_0, \mathbf{G}_1, \ldots, \mathbf{G}_p)$ *is also a time dependent adjustment set relative to* $(\mathbf{A}, Y)$ *in* $\mathcal{G}$ *and for all* $P \in \mathcal{M}(\mathcal{G})$,

$$\sigma^2_{\mathbf{a},\mathbf{G},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{G}}(P) \geq 0 \quad \textit{and} \quad \sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P) - \sigma^2_{\Delta,\mathbf{G}}(P) \geq 0.$$

In the Appendix we provide formulas for $\sigma^2_{\mathbf{a},\mathbf{G},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{G}}(P)$ *and* $\sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P) - \sigma^2_{\Delta,\mathbf{G}}(P)$.

It is interesting to contrast the requirements in Lemma 12 to those in Lemma 5. Lemma 12 requires the conditional independence (14) for the intermediate covariates $\mathbf{G}_j$. To get an intuition for why this requirement arises consider the case $p = 1$. For the covariate adjustment set $\mathbf{Z} = (\mathbf{Z}_0, \mathbf{Z}_1)$, where $\mathbf{Z}_0 = (\mathbf{G}_0, \mathbf{B}_0)$ and $\mathbf{Z}_1 = (\mathbf{G}_1, \mathbf{B}_1)$, the functional of interest is

$$\chi_{(a_0,a_1)}(P; \mathcal{G}) = E_P\left[E_P\left[E_P\left[Y \mid A_0 = a_0, A_1 = a_1, \mathbf{G}_0, \mathbf{B}_0, \mathbf{G}_1, \mathbf{B}_1\right] \mid A_0 = a_0, \mathbf{G}_0, \mathbf{B}_0\right]\right].$$

We can regard the problem of estimating the right hand side of the last display as a sequence of two estimation problems. The first is to estimate

$$E_P\left[E_P\left[Y \mid A_0 = a_0, A_1 = a_1, \mathbf{G}_0, \mathbf{B}_0, \mathbf{G}_1, \mathbf{B}_1\right] \mid A_0 = a_0, \mathbf{G}_0, \mathbf{B}_0\right].$$

Within levels of $\mathbf{G}_0, \mathbf{B}_0$ and $A_0 = a_0$ this problem is identical to the estimation of the interventional mean for a point exposure treatment that sets A_1 to a_1 with the time independent adjustment set $(\mathbf{G}_1, \mathbf{B}_1)$. By Lemma 5 we know that if $Y \perp\!\!\!\perp \mathbf{B}_1 \mid \mathbf{G}_1, \mathbf{G}_1, \mathbf{B}_0, A_0, A_1$ then $\mathbf{G}_1$ is also an adjustment set and is more efficient than $(\mathbf{G}_1, \mathbf{B}_1)$.

To understand the second estimation problem notice that when $Y \perp\!\!\!\perp \mathbf{B}_0, \mathbf{B}_1 \mid \mathbf{G}_1, \mathbf{G}_1, A_0, A_1$, the last display is equal to

$$E_P\left[E_P\left[Y \mid A_0 = a_0, A_1 = a_1, \mathbf{G}_0, \mathbf{G}_1\right] \mid A_0 = a_0, \mathbf{G}_0, \mathbf{B}_0\right] = E_P\left[h(\mathbf{G}_0, \mathbf{G}_1) \mid A_0 = a_0, \mathbf{G}_0, \mathbf{B}_0\right].$$

where $h(\mathbf{G}_0, \mathbf{G}_1) = E_P\left[Y \mid A_0 = a_0, A_1 = a_1, \mathbf{G}_0, \mathbf{G}_1\right]$. Then, the second estimation problem is to estimate

$$E_P\left[E_P\left[h(\mathbf{G}_0, \mathbf{G}_1) \mid A_0 = a_0, \mathbf{G}_0, \mathbf{B}_0\right]\right]. \tag{15}$$

The last display can be viewed as another interventional mean, now for a point exposure treatment that sets A_0 to a_0, with time independent adjustment set $(\mathbf{G}_0, \mathbf{B}_0)$, but with pseudo-outcome $h(\mathbf{G}_0, \mathbf{G}_1)$. Condition (14) with $p = 1$ is the condition that $\mathbf{G}_1 \perp\!\!\!\perp \mathbf{B}_0 \mid A_0, \mathbf{G}_0$. This yields $h(\mathbf{G}_0, \mathbf{G}_1) \perp\!\!\!\perp \mathbf{B}_0 \mid A_0, \mathbf{G}_0$. Then again by Lemma 5 we obtain that $\mathbf{G}_0$ is an adjustment set for the interventional mean in display (15), which is more efficient than $(\mathbf{G}_0, \mathbf{B}_0)$. Combining both results we conclude that $(\mathbf{G}_0, \mathbf{G}_1)$ is more efficient than $[(\mathbf{G}_0, \mathbf{B}_0), (\mathbf{G}_1, \mathbf{B}_1)]$.

We now have the following corollary to Lemmas 11 and 12.

Theorem 13 *Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$, let $\mathbf{A} = (A_0, \ldots, A_p)$ be a topologically ordered vertex set in $\mathbf{V}$ disjoint with $Y \in \mathbf{V}$. Assume $A_j, j = 0, \ldots, p$, correspond to finite valued random variables. Suppose*

$$\mathbf{B} = (\mathbf{B}_0, \ldots, \mathbf{B}_p) \subset \mathbf{V} \backslash \{\mathbf{A}, Y\}$$

and

$$\mathbf{G} = (\mathbf{G}_0, \ldots, \mathbf{G}_p) \subset \mathbf{V} \backslash \{\mathbf{A}, Y\}$$

are two time dependent adjustment sets relative to $(\mathbf{A}, Y)$ in $\mathcal{G}$. Suppose that

$$A_j \perp\!\!\!\perp_{\mathcal{G}} \left[\overline{\mathbf{G}}_j \backslash \overline{\mathbf{B}}_j\right] \mid \overline{\mathbf{B}}_j, \overline{\mathbf{A}}_{j-1} \text{ for } j = 0, \ldots, p$$

$$Y \perp\!\!\!\perp_{\mathcal{G}} [\mathbf{B} \backslash \mathbf{G}] \mid \mathbf{G}, \mathbf{A}$$

and

$$\mathbf{G}_j \perp\!\!\!\perp_{\mathcal{G}} \left[\overline{\mathbf{B}}_{j-1} \backslash \overline{\mathbf{G}}_{j-1}\right] \mid \overline{\mathbf{G}}_{j-1}, \overline{\mathbf{A}}_{j-1} \text{ for } j = 1, \ldots, p.$$

Then for all $P \in \mathcal{M}(\mathcal{G})$,

$$\sigma^2_{\mathbf{a},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{G}}(P) \geq 0 \quad \text{and} \quad \sigma^2_{\Delta,\mathbf{B}}(P) - \sigma^2_{\Delta,\mathbf{G}}(P) \geq 0.$$

Proof Write $\sigma^2_{\mathbf{a},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{G}}(P) = \sigma^2_{\mathbf{a},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{B}\cup(\mathbf{G}\backslash\mathbf{B})}(P) + \sigma^2_{\mathbf{a},\mathbf{G}\cup(\mathbf{B}\backslash\mathbf{G})}(P) - \sigma^2_{\mathbf{a},\mathbf{G}}(P)$ and apply Lemmas 11 and 12. ■

As indicated earlier, optimal adjustment sets always exist for point interventions. In contrast, for joint interventions, even though time dependent adjustment sets always exist, there exist DAGs with no optimal time dependent adjustment set. Moreover, even when an optimal time independent adjustment set exists, this set is not necessarily uniformly optimal among all adjustment sets. The following example illustrates these two points, as well as the application of Theorem 13.

Example 4 *For the DAG in Figure 4, Table 1 lists all valid time dependent adjustment sets relative to $(\mathbf{A},Y)$ for $\mathbf{A} = (A_0, A_1)$. These can be found applying the sufficient criteria in Pearl and Robins (1995). To see that no other valid time dependent adjustment set can exist, write*

$$E[Y_{\mathbf{a}}] = E_P\left[E_P\left[Y \mid A_0 = a_0, A_1 = a_1, H\right]\right], \tag{16}$$

which follows by recalling that $\mathbf{Z}_0 = \text{pa}_{\mathcal{G}}(A_0)$ *and* $\mathbf{Z}_1 = \text{pa}_{\mathcal{G}}(A_1) \setminus \text{pa}_{\mathcal{G}}(A_1)$ *is a valid adjustment set, as indicated in Section 3.1. Because the right hand side of* (16) *is not equal to*

$$E_P\left[E_P\left[Y \mid A_0 = a_0, A_1 = a_1\right]\right],$$

it follows that $\mathbf{Z}_0 = \emptyset$ *and* $\mathbf{Z}_1 = \emptyset$ *is not a valid time dependent adjustment set. Because the adjustment sets 1-11 in Table 1 exhaust all possible adjustment sets such that* $\mathbf{Z}_0$ *excludes* R *and* Q*, then no other valid adjustment set can exist since* R *and* Q *are in the causal path between* A_0 *and* Y *and therefore cannot be included in* $\mathbf{Z}_0$*.*

The last column of Table 1 indicates, for every adjustment set, another dominating adjustment set, in the sense that the dominating one results in an NP-$\mathbf{Z}$ *estimator that has smaller asymptotic variance for all* $P \in \mathcal{M}(\mathcal{G})$*. These dominating adjustment sets are found applying Theorem 13. We note however that* $\mathbf{Z}^*$ *in row 1 is superior to* $\mathbf{Z}^{**}$ *in row 8 for some* $P \in \mathcal{M}(\mathcal{G})$ *but* $\mathbf{Z}^{**}$ *is superior to* $\mathbf{Z}^*$ *for another* $P' \in \mathcal{M}(\mathcal{G})$*. Intuitively, when the association encoded in the red arrow is weak but the associations encoded in the blue arrows are strong, then* $\mathbf{Z}^{**}$ *in row 8 is preferable to* $\mathbf{Z}^*$ *in row 1. In contrast, when the association encoded in the red arrow is strong but the associations encoded in the blue arrows are weak, then* $\mathbf{Z}^*$ *if preferable to* $\mathbf{Z}^{**}$*. For instance, when all variables are binary there exists* $P \in \mathcal{M}(\mathcal{G})$ *such that* $\sigma^2_{\mathbf{a},\mathbf{Z}^*}(P)/\sigma^2_{\mathbf{a},\mathbf{Z}^{**}}(P) = 0.675$ *and another law* $P' \in \mathcal{M}(\mathcal{G})$ *such that* $\sigma^2_{\mathbf{a},\mathbf{Z}^*}(P')/\sigma^2_{\mathbf{a},\mathbf{Z}^{**}}(P') = 1.08$ *for* $\mathbf{a} = (1,1)$*. See the* `R` *scripts available at* `https://github.com/esmucler/optimal_adjustment`*. We note also that the adjustment set in row 11, namely* $\mathbf{Z}_0 = \{H\}$ *and* $\mathbf{Z}_1 = \emptyset$ *is the unique time independent adjustment set, and hence optimal among time independent adjustment sets. Nevertheless, it is dominated by the time dependent adjustment set in row 8, thus proving that optimal time independent adjustment sets need not be optimal in the class of all adjustment sets.*

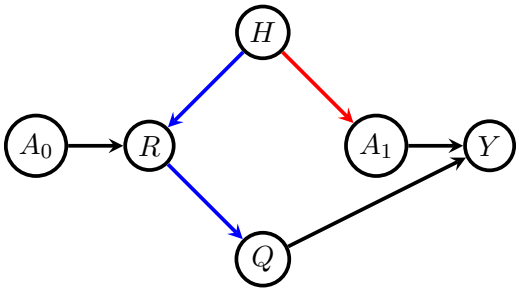

Figure 4: An example in which no optimal time dependent adjustment set exists.

Adjustment set	$\mathbf{Z}_0$	$\mathbf{Z}_1$	Dominating adjustment set
1	$\emptyset$	Q	-
2	$\emptyset$	R	1
3	$\emptyset$	H	1
4	$\emptyset$	$\{Q, R\}$	1
5	$\emptyset$	$\{Q, H\}$	1
6	$\emptyset$	$\{R, H\}$	1
7	$\emptyset$	$\{Q, R, H\}$	1
8	H	Q	-
9	H	R	8
10	H	$\{R, Q\}$	8
11	H	$\emptyset$	8

Table 1: List of all possible time dependent adjustment sets for the DAG in Figure 4.

An interesting open question is to characterize the class of DAGs for which there exists an optimal time dependent adjustment set.

5.3 Non-existence of Uniformly Optimal Covariate Adjustment Sets in Non-parametric Causal Graphical Models with Latent Variables

Consider now the situation in which some vertices of the DAG are not observable, but some adjustment sets are observable. A natural question is whether one can find an optimal adjustment set among the observable ones. Without restricting the topology of the DAG, the answer is negative as the following example illustrates. For linear causal graphical models and treatments effects estimated by OLS, Henckel et al. (2019) showed that it is possible that no uniformly optimal adjustment set exists among observable adjustment sets. In the following example we show the same negative result holds for non-linear causal graphical models and NP-**O** estimators, where by an NP-**O** estimator we mean the NP-**Z** estimator that uses $\mathbf{Z} = \mathbf{O}$.

An interesting open problem is the characterization of settings in which, $\mathbf{O}(\mathbf{A}, Y, \mathcal{G})$ is not observed but an optimal observable adjustment set exists.

Example 5 *Suppose that in the DAG in Figure 5, U is the only unobserved variable. Then, $\mathbf{Z}^* = \emptyset$, $\mathbf{Z}^{**}=\{Z_1, Z_2\}$ and $\mathbf{Z}^{***} = \{Z_1\}$ are all observable adjustment sets for (A, Y) relative to the DAG. Using Lemma 5, it is easy to show that $\mathbf{Z}^*$ is uniformly better than $\mathbf{Z}^{***}$. However, $\mathbf{Z}^*$ is better than $\mathbf{Z}^{**}$ if the associations encoded in the blue edges are strong and the associations encoded in the red edges are weak, but $\mathbf{Z}^{**}$ is better than $\mathbf{Z}^*$ if the blue edges are weak and red ones are strong. In fact, when all variables are binary there exists $P \in \mathcal{M}(\mathcal{G})$ such that $\sigma^2_{1,\mathbf{Z}^*}(P)/\sigma^2_{1,\mathbf{Z}^{**}}(P) = 0.04$ and another $P' \in \mathcal{M}(\mathcal{G})$ such that $\sigma^2_{1,\mathbf{Z}^*}(P')/\sigma^2_{1,\mathbf{Z}^{**}}(P') = 1.44$. See the `R` scripts available at `https://github.com/esmucler/optimal_adjustment`.*

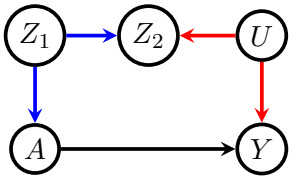

Figure 5: An example with a latent variable U and observable covariate adjustment sets but with no optimal adjustment set.

6. Estimation of Point Intervention Causal Effects Exploiting the Assumptions of the Bayesian Network

For a point intervention $\mathbf{A} = A$, NP-$\mathbf{O}$ estimators of the individual interventional means and their contrasts, even though efficient among NP-$\mathbf{Z}$ estimators, ignore the conditional independence assumptions encoded in the causal graphical model about the data generating law P. These assumptions may carry information about the parameters of interest. For instance, consider the Bayesian Network represented by the DAG $\mathcal{G}$ of Figure 6. Under model $\mathcal{M}(\mathcal{G})$, the components O_1 and O_2 of the adjustment set $\mathbf{O} = \{O_1, O_2\}$ are marginally independent. This independence carries information about $\chi_a(P;\mathcal{G}) = E_P[E_p[Y|A=a,O_1,O_2]]$ because the joint distribution of O_1, O_2 is not ancillary for $\chi_a(P;\mathcal{G})$. Specifically,

$$\begin{aligned} E_P[E_p[Y|A=a,O_1,O_2]] &= \int\int\int yp(y|a,o_1,o_2)\,p(o_1,o_2)\,dydo_1do_2 \\ &= \int\int\int yp(y|a,o_1,o_2)\,p(o_1)\,p(o_2)\,dydo_1do_2 \end{aligned}$$

and the last equality is true only under $\mathcal{M}(\mathcal{G})$.

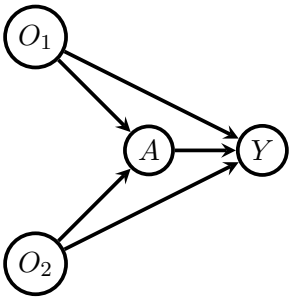

Figure 6: A DAG where the NP-$\mathbf{O}$ estimator is inefficient.

Applying Algorithm 2 of Section 6.4, it is easy show that the semiparametric Cramer-Rao bound under the Bayesian Network of Figure 6 is equal to the variance of the random variable $\chi^1_{P,a,eff}(A,Y,\mathbf{O};\mathcal{G}) = \psi_{P,a}(\mathbf{Z};\mathcal{G}) - \Delta_P(\mathbf{O})$ where $\mathbf{O} = (O_1, O_2)$, $\psi_{P,a}(\mathbf{O};\mathcal{G})$ is the influence function of the NP-$\mathbf{O}$ estimator defined in (4) and

$$\Delta_P(\mathbf{O}) \equiv b_a(\mathbf{O};P) - E_P[b_a(\mathbf{O};P)|O_1] - E_P[b_a(\mathbf{O};P)|O_2] + E_P[b_a(\mathbf{O};P)]$$

with $b_a(\mathbf{O};P) \equiv E_P[Y|A=a,\mathbf{O}]$. Furthermore, we show in Lemma 23 in the Appendix that if $P_\alpha \in \mathcal{M}(\mathcal{G})$ is such that the following hold (i) $b_a(\mathbf{O};P_\alpha) = O_1 + O_2 + \alpha O_1 O_2$, (ii) $E_{P_\alpha}(O_1) = E_{P_\alpha}(O_2) = 0$, (iii) $E_{P_\alpha}\left(O_1^2\right) = E_{P_\alpha}\left(O_2^2\right) = 1$, (iv) there exists a fixed $C > 0$ independent of α such that $var_{P_\alpha}(Y \mid A=a,\mathbf{O}) \leq C$ and $\pi_a(\mathbf{O}_{min};P_\alpha) \geq 1/C$, then $\Delta_{P_\alpha}(\mathbf{O}) = \alpha O_1 O_2$ and

$$\frac{var_{P_\alpha}\left[\psi_{P_\alpha,a}(\mathbf{O};\mathcal{G})\right]}{var_{P_\alpha}\left[\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})\right]} \underset{|\alpha|\to\infty}{\to} \infty.$$

This illustrates the point that the NP-**O** estimator may ignore a substantial fraction of the information about $\chi_a(P;\mathcal{G})$ encoded in the Bayesian Network.

Independencies among variables in the adjustment set are not the only carriers of information about $\chi_a(P;\mathcal{G})$ in a Bayesian Network. For instance, consider the model represented by the DAG in Figure 7 in which A is randomized but a variable M that mediates all the effect of A on Y is measured. Then

$$\chi_a(P;\mathcal{G}) = E_p[Y|A=a] = \int\int yp(y,m|a)\,dydm = \int\int yp(y|m)\,p(m|a)\,dydm$$

and the last equality holds due to the Markov chain structure encoded in the model. In this example, the empirical mean of Y given $A=a$, can be viewed as the NP-**O** estimators where $\mathbf{O} = \emptyset$. However this estimator does not attain the semiparametric Cramer-Rao bound, because it does not exploit the Markov chain structure encoded in the graph.

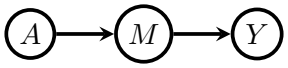

Figure 7: A DAG where the NP-**O** estimator is inefficient.

As a third example, consider the Bayesian Network represented by the DAG in Figure 8. Under this model O is the unique, and hence optimal, covariate adjustment set. Nev-

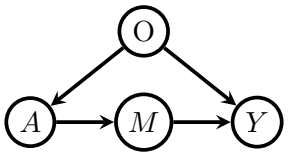

Figure 8: The front-door graph.

ertheless, under the model, $\chi_a(P;\mathcal{G}) = E_P[E_p[Y|A=a,O]]$ is also equal to the so-called front-door functional

$$\beta(P) \equiv \int y\left\{\int p(m|a)\left[\sum_{a'} p\left(y|m,a'\right)p\left(a'\right)\right]dm\right\}dy. \tag{17}$$

See Pearl (2000). This example has been used in the literature to illustrate an instance, when O is unobserved, in which identifiability of $\chi_a(P,\mathcal{G})$ is possible even if no covariate adjustment is possible. Under regularity conditions, the non-parametric estimator of

$\beta(P)$, based on estimating the right hand side of (17) replacing all densities by smooth non-parametric estimators of them, provides yet another regular and asymptotically linear estimator of $\chi_a(P;\mathcal{G})$. In fact, in Example 7 we argue that neither estimator attains the semiparametric Cramer-Rao bound under the model (see Hayashi and Kuroki (2014) for a related discussion in causal linear models). The one-step estimation technique described in Section 6.1 can be used to obtain a regular and asymptotically linear estimator of $\chi_a(P;\mathcal{G})$ which, under regularity conditions, attains the semiparametric Cramer-Rao Bound under $\mathcal{M}(G)$.

In Section 6.1 we provide estimators of $\chi_a(P;\mathcal{G})$ and $\Delta(P,\mathcal{G})$ that, under regularity conditions, converge to a zero mean normal distribution with a variance that attains the semiparametric Cramer-Rao bound for the parameter under the model $\mathcal{M}(\mathcal{G})$. In Section 6.2 we provide a sound and complete graphical algorithm that, given a DAG $\mathcal{G}$, decides whether or not under all laws P of $\mathcal{M}(\mathcal{G})$, the NP-**O** estimator is semiparametric efficient under the Bayesian Network $\mathcal{M}(\mathcal{G})$. Furthermore, in Section 6.4 we provide an algorithm that, when the NP-O estimator is not semiparametric efficient, returns a simplified formula for the efficient influence function, which facilitates the computation of an estimator that under regularity conditions attains the semiparametric Cramer-Rao bound.

Because estimation of causal effects under a DAG $\mathcal{G}$ is only meaningful when in $\mathcal{G}$ there exists at least one causal path between A and Y, from now on we will consider only inference about $\chi_a(P;\mathcal{G})$ under Bayesian Networks $\mathcal{M}(\mathcal{G})$ represented by such DAGs.

6.1 Semiparametric Efficient Estimation

The problem we are concerned with in this section is formalized as follows. We are interested in finding an estimator of the functionals $\chi_a(P;\mathcal{G}) \equiv E_P\left[E_P\left[Y|A=a,\mathrm{pa}_{\mathcal{G}}(A)\right]\right]$ and $\Delta(P;\mathcal{G}) \equiv \sum_{\mathbf{a}\in\mathcal{A}} c_{\mathbf{a}} E[Y_{\mathbf{a}}]$, with the smallest possible variance among all estimators that are regular and asymptotically linear under any $P \in \mathcal{M}(\mathcal{G})$. When not all the variables in $\mathcal{G}$ are discrete, the tangent space of model $\mathcal{M}(\mathcal{G})$ is not a subset of a Euclidean space. Consequently, in such case model $\mathcal{M}(\mathcal{G})$ is semiparametric. In fact, in Lemma 24 of the Appendix, we show that $\Lambda \equiv \oplus_{j=1}^{s}\Lambda_j$ where

$$\Lambda_j \equiv \left\{G \equiv g\left(V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right) \in L_2(P) : E_P\left[G|\mathrm{pa}_{\mathcal{G}}(V_j)\right] = 0\right\}.$$

and $\oplus$ stands for the sum of $L_2(P)-$orthogonal spaces. Thus, unless $\mathcal{G}$ is a complete DAG, Λ is a strict subset of $L_2^0(P)$, where $L_2^0(P) \equiv \left\{g \in L_2(P) : \int g dP = 0\right\}$.

Notice that by the linearity of the differentiation operation, if $\chi^1_{P,a}(\mathbf{V};\mathcal{G})$ denotes an influence function for $\chi_a(P;\mathcal{G})$ then $\Delta^1_P(\mathbf{V};\mathcal{G}) = \sum_{a\in\mathcal{A}} c_a \chi^1_{P,a}(\mathbf{V};\mathcal{G})$ is an influence function for $\Delta(P;\mathcal{G})$. Consequently, if $\Delta^1_{P,eff}(\mathbf{V};\mathcal{G})$ and $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$ denote the efficient influence functions of $\Delta(P;\mathcal{G})$ and $\chi_a(P;\mathcal{G})$, we have

$$\Delta^1_{P,eff}(\mathbf{V};\mathcal{G}) = \sum_{a\in\mathcal{A}} c_a \chi^1_{P,a,eff}(\mathbf{V};\mathcal{G}).$$

The next theorem provides an expression for $\chi^1_{P,a,eff}$. Let

$$J_{P,a,\mathcal{G}} \equiv \frac{I_a(A)\,Y}{P\left(A=a|\mathrm{pa}_{\mathcal{G}}(A)\right)},$$

$$\text{indir}\left(A,Y,\mathcal{G}\right) \equiv \{V_j \in \mathbf{V} : V_j \in \text{an}_{\mathcal{G}}\left(A\right) \setminus \{A\} \text{ and all causal paths between } V_j \text{ and } Y \text{ intersect } A\}$$

and

$$\text{irrel}\left(A,Y,\mathcal{G}\right) \equiv \text{indir}\left(A,Y,\mathcal{G}\right) \cup \text{an}_{\mathcal{G}}\left(Y\right)^c.$$

Note that $\text{indir}\left(A,Y,\mathcal{G}\right)$ is comprised by the nodes in $\mathbf{V}$ that, conditional on their parents, are instrumental variables for the causal effect of A on Y. Note also that $\text{irrel}(A,Y,\mathcal{G}) = [\text{an}_{\mathcal{G}_{\mathbf{V}\setminus\{A\}}}(Y)]^c$. See also Perković (2019).

The following theorem establishes that the efficient influence function of $\chi^1_{P,a,eff}\left(\mathbf{V};\mathcal{G}\right)$ does not depend on the variables in $\text{irrel}(A,Y,\mathcal{G})$.

Theorem 14 *Let $\mathcal{M}\left(\mathcal{G}\right)$ be the Bayesian Network represented by DAG $\mathcal{G}$ with vertex set $\mathbf{V}$. Assume Y and A are two distinct vertices. Then, the efficient influence function of $\chi_a\left(P;\mathcal{G}\right)$ at P under $\mathcal{M}\left(\mathcal{G}\right)$ is equal to*

$$\chi^1_{P,a,eff}\left(\mathbf{V};\mathcal{G}\right) = \sum_{j:V_j \notin [\text{irrel}(A,Y,\mathcal{G})\cup\{A\}]} \left\{E_P\left[J_{P,a,\mathcal{G}}|V_j, \text{pa}_{\mathcal{G}}\left(V_j\right)\right] - E_P\left[J_{P,a,\mathcal{G}}|\,\text{pa}_{\mathcal{G}}\left(V_j\right)\right]\right\}. \tag{18}$$

Furthermore, $\chi^1_{P,a,eff}\left(\mathbf{V};\mathcal{G}\right)$ depends on $\mathbf{V}$ only through $\mathbf{V}_{marg} \equiv \mathbf{V}\setminus \text{irrel}\left(A,Y,\mathcal{G}\right)$.

Our next results will establish that the variables in $\text{irrel}(A,Y,\mathcal{G})$ can be marginalized from the DAG without incurring in any loss of information about the parameter. Recall that for any DAG $\mathcal{G}$ with vertex set $\mathbf{V}$ and a subset of nodes $\mathbf{V}_{marg}$, $\mathcal{M}\left(\mathcal{G},\mathbf{V}_{\text{marg}}\right)$ denotes the marginal DAG model. See Section 2.1.

Definition 15 *Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$ and $\mathbf{V}_{marg} \subset \mathbf{V}$. For any $P \in \mathcal{M}\left(\mathcal{G}\right)$ let P_{marg} denote the marginal distribution of $\mathbf{V}_{marg}$ under P. Let $\mathcal{G}'$ be a DAG with vertex set $\mathbf{V}_{marg}$. Let A, Y be two distinct nodes such that $\{A,Y\} \subset \mathbf{V}_{marg}$. We say that $(\mathbf{V}_{marg}, \mathcal{G}')$ is sufficient for efficient estimation of $\chi_a\left(P;\mathcal{G}\right)$ relative to $(\mathbf{V},\mathcal{G})$ if for all $P \in \mathcal{M}\left(\mathcal{G}\right)$, the following conditions hold*

1. $\mathcal{M}\left(\mathcal{G},\mathbf{V}_{marg}\right) = \mathcal{M}\left(\mathcal{G}'\right)$,
2. $\chi_a\left(P;\mathcal{G}\right) = \chi_a\left(P_{marg};\mathcal{G}'\right)$,
3. $\mathbf{O}\left(A,Y,\mathcal{G}\right) = \mathbf{O}\left(A,Y,\mathcal{G}'\right)$,
4. $\psi_{P,a}\left[\mathbf{O}\left(A,Y,\mathcal{G}\right);\mathcal{G}\right] = \psi_{P_{marg},a}\left[\mathbf{O}\left(A,Y,\mathcal{G}'\right);\mathcal{G}'\right]$ *and*
5. $\chi^1_{P,a,eff}\left(\mathbf{V};\mathcal{G}\right) = \chi^1_{P_{marg},a,eff}\left(\mathbf{V}_{marg};\mathcal{G}'\right)$

If we find $(\mathbf{V}_{\text{marg}}, \mathcal{G}')$ that is sufficient for estimation of $\chi_a\left(P;\mathcal{G}\right)$ relative to $(\mathbf{V},\mathcal{G})$, then we do not incur in any loss of information about $\chi_a\left(P;\mathcal{G}\right)$ if we ignore the variables in $\mathbf{V}\setminus\mathbf{V}_{\text{marg}}$ and assume that $\mathbf{V}_{\text{marg}}$ follows a Bayesian Network $\mathcal{M}\left(\mathcal{G}'\right)$ for the DAG $\mathcal{G}'$. Furthermore, since $\mathcal{G}'$ preserves the optimal adjustment set then the NP-**O** estimator of $\chi_a\left(P;\mathcal{G}\right)$ is the same as the NP-**O** estimator of $\chi_a\left(P_{\text{marg}};\mathcal{G}'\right)$. Since by condition 5) of the preceding definition the efficiency bound for $\chi_a\left(P;\mathcal{G}\right)$ under $\mathcal{M}\left(\mathcal{G}\right)$ is the same as

the efficiency bound for $\chi_a(P_{\text{marg}}; \mathcal{G}')$ under $\mathcal{M}(\mathcal{G}')$, then for studying the loss of efficiency incurred by using the NP-**O** estimator we can pretend that the available variables are $\mathbf{V}_{\text{marg}}$ and that the problem is to estimate $\chi_a(P_{\text{marg}}; \mathcal{G}')$ under the Bayesian Network $\mathcal{M}(\mathcal{G}')$.

The next lemma implies that $\mathbf{V}_{\text{marg}} = \mathbf{V} \setminus \text{irrel}(A, Y, \mathcal{G})$ and $\mathcal{G}'$ equal to the output of Algorithm 1 below satisfies the preceding definition.

Lemma 16 *Let $\mathcal{G}$ and $\mathcal{G}'$ be the input and output DAGs of Algorithm 1. Let $\mathbf{V}$ and $\mathbf{V}_{marg}$ be the vertex sets of $\mathcal{G}$ and $\mathcal{G}'$ respectively. Then $(\mathbf{V}_{marg}, \mathcal{G}')$ is sufficient for efficient estimation of $\chi_a(P; \mathcal{G})$ relative to $(\mathbf{V}, \mathcal{G})$.*

Algorithm 1: DAG pruning procedure to remove irrelevant nodes

input : DAG $\mathcal{G}$ with nodes $\mathbf{V}$ and two distinct nodes $A, Y \in \mathbf{V}$
output: A new DAG $\mathcal{G}'$ with vertex set $\mathbf{V}_{marg} = \mathbf{V} \setminus \text{irrel}(A, Y, \mathcal{G})$ such that $(\mathbf{V}_{marg}, \mathcal{G}')$ is sufficient for efficient estimation of $\chi_a(P; \mathcal{G})$ relative to $(\mathbf{V}, \mathcal{G})$.
procedure `prune`$(A, Y, \mathcal{G})$
  $\mathcal{G}' = \mathcal{G}_{\text{an}_{\mathcal{G}}(Y)}$
  $I_1, \ldots, I_L =$ `topological_sort` $(\text{indir}(A, Y, \mathcal{G}'), \mathcal{G}')$
  for $j = L, L-1, \ldots, 1$ **do**
    $\mathcal{G}' = \tau(\mathcal{G}', I_j)$
  return $\mathcal{G}'$;

The output $\mathcal{G}'$ of Algorithm 1 is obtained as the result of first deleting the edges and vertices in $\text{an}_{\mathcal{G}}(Y)^c$ and subsequently removing, sequentially by a latent projection operation, each node in $\text{indir}(A, Y, \mathcal{G})$. For the definition of the latent projection operation $\tau(\mathcal{G}, V)$ see Section 2.1. In general the set of possible marginal distributions for a subvector $\mathbf{V}'$ of a vector $\mathbf{V}$ following a Bayesian Network $\mathcal{M}(\mathcal{G})$ is not necessarily a Bayesian Network $\mathcal{M}(\mathcal{G}')$ for some DAG $\mathcal{G}'$ with vertex set $\mathbf{V}'$ (Evans, 2016). However, because of the specific structure of the set $\text{irrel}(A, Y, \mathcal{G})$, the set of possible marginal distributions for $\mathbf{V}' = \mathbf{V}_{marg} = \mathbf{V} \setminus \text{irrel}(A, Y, \mathcal{G})$ is the Bayesian Network $\mathcal{M}(\mathcal{G}')$, where $\mathcal{G}'$ is the output of Algorithm 1. Algorithm 1 assumes the availability of a subroutine `topological_sort` to topologically sort a set of nodes relative to a DAG $\mathcal{G}$. One such subroutine is Kahn's algorithm (Kahn, 1962), which is known to have worst case complexity $\mathcal{O}(|\mathbf{V}| + |\mathbf{E}|)$.

Example 6 *We illustrate an application of Algorithm 1 with the graph in Figure 9 (a). In this graph, $\text{an}^c_{\mathcal{G}}(Y) = \{D_1, D_2\}$ and $\text{indir}(A, Y, \mathcal{G}) = \{I_1, I_2, I_3\}$, which are already topologically ordered. The first step of the procedure* `prune` *removes all nodes in $\text{an}^c_{\mathcal{G}}(Y)$ and all edges into them from the graph. The next step of the procedure marginalizes over I_3. The graph in Figure 9 (b) is the result of this marginalization applied to the graph with $\{D_1, D_2\}$ already removed. The next step marginalizes over I_2 and the resulting graph is shown in Figure 9 (c). The last step marginalizes over I_1 yielding the graph in Figure 9 (d) which is the output of Algorithm 1.*

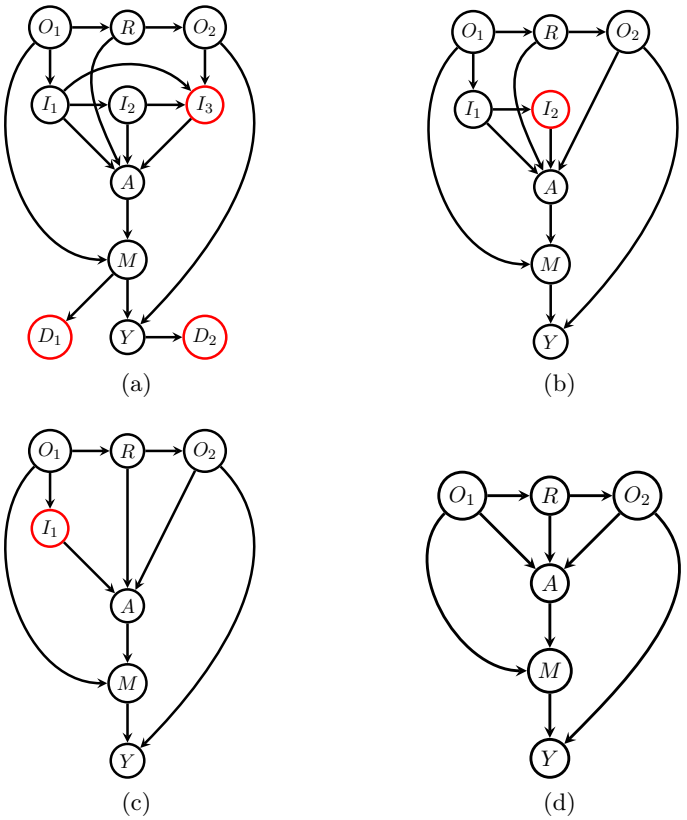

Figure 9: An example of an application of Algorithm 1.

Lemma 16 is proven in the Appendix by invoking the following important result.

Proposition 17 *Let $\mathcal{M}$ be a semiparametric model for the law of a random vector $\mathbf{V}$. Let $\mathbf{V}'$ be a subvector of $\mathbf{V}$. Let $\mathcal{M}'$ be the model for the law of $\mathbf{V}'$ induced by model $\mathcal{M}$, that is, $\mathcal{M}'$ is the collection of laws for $\mathbf{V}'$ such that for every $P' \in \mathcal{M}'$ there exists a law P for $\mathbf{V}$ with P' being the marginal of P over $\mathbf{V}'$. Let $\chi(P)$ be a regular parameter in model $\mathcal{M}$ with efficient influence function at $P \in \mathcal{M}$ equal to $\chi^1_{P,eff}$. Suppose $\chi^1_{P,eff}$ depends on $\mathbf{V}$ only through $\mathbf{V}'$. Let P' be the marginal law of P over $\mathbf{V}'$. Suppose $\chi(P)$ depends on P only through P'. Define $\nu(P') \equiv \chi(P)$. Let $\nu^1_{P',eff}$ be the efficient influence function of $\nu(P')$ in model $\mathcal{M}'$ at $P' \in \mathcal{M}'$. Then, given $P' \in \mathcal{M}'$ it holds that $\chi^1_{P,eff} = \nu^1_{P',eff}$ for every $P \in \mathcal{M}$ with marginal law P'.*

In light of Lemma 16 and Proposition 17, from now on without loss of generality we will assume that $\text{irrel}(A, Y, \mathcal{G}) = \emptyset$. This assumption implies that we can partition the node set $\mathbf{V}$ of $\mathcal{G}$ as $\mathbf{M} \cup \mathbf{W} \cup \{A, Y\}$ where the vertices in $\mathbf{M}$ intersect at least one causal path between A and Y, that is, $\mathbf{M}$ is the set of mediators in the causal pathways between A and Y, and $\mathbf{W}$ are non-descendants of A. We can therefore sort topologically $\mathbf{V}$ as $(W_1, \ldots, W_J, A, M_1, \ldots, M_K, Y)$. The set $\mathbf{O}(A, Y, \mathcal{G}) \equiv \mathbf{O} \equiv (O_1, \ldots, O_T)$, where $(O_1, \ldots, O_T)$ is sorted topologically, is included in $\mathbf{W}$. Throughout $T = 0$ if $\mathbf{O}(A, Y, \mathcal{G}) = \emptyset$.

The following Theorem establishes another expression for $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$. Subsequently we exploit this expression to describe a strategy to construct a semiparametric efficient estimator of $\chi_a(P;\mathcal{G})$. Let

$$T_{P,a,\mathcal{G}} \equiv \frac{I_a(A)\,Y}{\pi_a(\mathbf{O};P)}.$$

In what follows we use the conventions $\sum_{k=1}^{0}\cdot \equiv 0, \sum_{j=1}^{0}\cdot \equiv 0, \sum_{j=2}^{1}\cdot \equiv 0$.

Theorem 18 *Let $\mathcal{M}(\mathcal{G})$ be the Bayesian Network represented by DAG $\mathcal{G}$ with vertex set $\mathbf{V}$. Assume Y and A are single disjoint vertices. Assume* $\text{irrel}(A,Y,\mathcal{G}) = \emptyset$. *Then the efficient influence function of $\chi_a(P;\mathcal{G})$ at P under $\mathcal{M}(\mathcal{G})$ is equal to*

$$\begin{aligned}\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G}) &= E_P\left[T_{P,a,\mathcal{G}}|Y,\text{pa}_{\mathcal{G}}(Y)\right] - E_P\left[T_{P,a,\mathcal{G}}|\,\text{pa}_{\mathcal{G}}(Y)\right] \\ &+\sum_{k=1}^{K}\left\{E_P\left[T_{P,a,\mathcal{G}}|M_k,\text{pa}_{\mathcal{G}}(M_k)\right] - E_P\left[T_{P,a,\mathcal{G}}|\,\text{pa}_{\mathcal{G}}(M_k)\right]\right\} \\ &+\sum_{j=1}^{J}\left\{E_P\left[b_a(\mathbf{O};P)|W_j,\text{pa}_{\mathcal{G}}(W_j)\right] - E_P\left[b_a(\mathbf{O};P)|\,\text{pa}_{\mathcal{G}}(W_j)\right]\right\}\end{aligned} \tag{19}$$

where $E_P\left[b_a(\mathbf{O};P)|\,\text{pa}_{\mathcal{G}}(W_1)\right] = \chi_a(P;\mathcal{G})$, because $\text{pa}_{\mathcal{G}}(W_1) = \emptyset$.

Example 7 *Consider again the DAG in Figure 8. In this example, $\mathbf{O} = \mathbf{O}_{min} = \{O\} \equiv \{W_1\}$, $J = T = 1$, $\mathbf{M} = \{M\}$ and $K = 1$. A quick calculation shows that equation (19) applied to this example yields*

$$\begin{aligned}\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G}) = b_a(O;P) - \chi_a(\mathcal{G};P) + YE_P\left[\frac{I_a(A)}{\pi_a(O;P)} \mid Y,M\right] - E_P\left[\frac{I_a(A)Y}{\pi_a(O;P)} \mid M\right] \\ + I_a(A)E_P\left[\frac{Y}{\pi_a(O;P)} \mid A,M\right] - I_a(A)E_P\left[\frac{Y}{\pi_a(O;P)} \mid A\right].\end{aligned}$$

Notice that for any law that is faithful to the DAG, $b_a(O;P)$ is a non-trivial function of O and $E_P\left[I_a(A)Y\pi_a^{-1}(O;P) \mid M\right]$ is a non-trivial function of M. Since these terms cannot cancel out with any of the remaining terms on the right hand side of the preceding display, we conclude that $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$ is a non-trivial function of both O and M. This shows that the NP-$\mathbf{O}$ estimator cannot be globally efficient in model $\mathcal{M}(\mathcal{G})$ because the NP-$\mathbf{O}$ estimator does not depend on M. It also demonstrates the point announced earlier, that the non-parametric estimator of the front-door formula (17) is also not globally efficient, because this estimator does not depend on the variable O.

The expression for $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$ in Theorem 18 can be used to compute the following one-step estimator (Van der Vaart, 2000; van der Vaart, 2014),

$$\widehat{\chi}_{one-step,a} \equiv \widehat{\chi}_a(P;\mathcal{G}) + \mathbb{P}_n\left[\widehat{\chi}^1_{P,a,eff}(\mathbf{V};\mathcal{G})\right]$$

where if $J = 0$, $\widehat{\chi}_a(P;\mathcal{G}) = \mathbb{P}_n[YI_a(A)]\{\mathbb{P}_n[I_a(A)]\}^{-1}$ and

$$\mathbb{P}_n\left[\widehat{\chi}^1_{P,a,eff}(\mathbf{V};\mathcal{G})\right] = \sum_{k=1}^{K}\mathbb{P}_n\left\{\widehat{E}\left[T_{P,a,\mathcal{G}}|M_k,\text{pa}_{\mathcal{G}}(M_k)\right] - \widehat{E}\left[T_{P,a,\mathcal{G}}|\,\text{pa}_{\mathcal{G}}(M_k)\right]\right\}$$

and if $J \geq 1$,

$$\begin{aligned}
\widehat{\chi}_{one-step,a} &\equiv \widehat{\chi}_a(P;\mathcal{G}) + \mathbb{P}_n\left[\widehat{\chi}^1_{P,a,eff}(\mathbf{V};\mathcal{G})\right] \\
&= \mathbb{P}_n\left\{\widehat{E}\left[T_{P,a,\mathcal{G}}|Y,\mathrm{pa}_{\mathcal{G}}(Y)\right] - \widehat{E}\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(Y)\right]\right\} \\
&\quad + \sum_{k=1}^{K}\mathbb{P}_n\left\{\widehat{E}\left[T_{P,a,\mathcal{G}}|M_k,\mathrm{pa}_{\mathcal{G}}(M_k)\right] - \widehat{E}\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(M_k)\right]\right\} \\
&\quad + \sum_{j=2}^{J}\mathbb{P}_n\left\{\widehat{E}\left[b_a(\mathbf{O};P)|W_j,\mathrm{pa}_{\mathcal{G}}(W_j)\right] - \widehat{E}\left[b_a(\mathbf{O};P)|\,\mathrm{pa}_{\mathcal{G}}(W_j)\right\}\right] \\
&\quad + \mathbb{P}_n\left\{\widehat{E}\left[b_a(\mathbf{O};P)|W_1,\mathrm{pa}_{\mathcal{G}}(W_1)\right]\right\}
\end{aligned}$$

and where $\widehat{E}(\cdot|\cdot)$ are non-parametric estimators of the relevant conditional expectations and $\mathbb{P}_n$ is the empirical mean operator. Specifically $\widehat{E}\left[T_{P,a,\mathcal{G}}|M_k,\mathrm{pa}_{\mathcal{G}}(M_k)\right]$ is constructed by first computing a non-parametric estimator $\widehat{\pi}_a(\mathbf{O})$ of $\pi_a(\mathbf{O};P)$ and subsequently computing a non-parametric regression estimator of the pseudo-outcomes $\widehat{T}_i$, where $\widehat{T}_i = Y_i I_a(A_i)/\widehat{\pi}_a(\mathbf{O}_i)$, on the covariates $M_{k,i},\mathrm{pa}_{\mathcal{G}}(M_{k,i})$ $i=1,\ldots,n$. Likewise, $\widehat{E}\left[b_a(\mathbf{O};P)|W_j,\mathrm{pa}_{\mathcal{G}}(W_j)\right]$ is constructed by first computing a non-parametric estimator regression estimator $\widehat{b}_a(\mathbf{O})$ of the mean of Y given $\mathbf{O}$ and $A=a$ and subsequently computing a non-parametric regression estimator of the pseudo-outcomes $\widehat{b}_a(\mathbf{O}_i)$ on the covariates $W_{j,i},\mathrm{pa}_{\mathcal{G}}(W_{j,i})$ $i=1,\ldots,n$. The estimators of the expectations that condition only on the parent set of a node are computed similarly.

Under regularity conditions, which include restrictions on some measure (for example, the metric entropy) of the complexity of the ambient function space of the conditional expectations appearing in the expression for $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$, and for particular choices of the non-parametric estimators of these conditional expectations, the one-step estimator $\widehat{\chi}_{one-step,a}$ is regular and asymptotically linear with influence function equal to $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$ (van der Vaart, 2014) and therefore it attains the semiparametric variance bound for $\chi_a(P;\mathcal{G})$ under model $\mathcal{M}(\mathcal{G})$. Consequently, $\widehat{\Delta} = \sum_{a\in\mathcal{A}} c_a\widehat{\chi}_{one-step,a}$ has influence function $\sum_{a\in\mathcal{A}} c_a\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$ and therefore it attains the semiparametric variance bound for $\Delta(P;\mathcal{G})$ under model $\mathcal{M}(\mathcal{G})$.

6.2 A Sound and Complete Algorithm to Check if the NP-O Estimator is Semiparametric Efficient

It turns out that for special configurations of $\mathcal{G}$, the formula (19) simplifies in that

$$\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G}) = \psi_{P,a}\left[\mathbf{O}(A,Y,\mathcal{G});\mathcal{G}\right] \text{ for all } P\in\mathcal{M}(\mathcal{G}), \tag{20}$$

where $\psi_{P,a}\left[\mathbf{O}(A,Y,\mathcal{G});\mathcal{G}\right]$ is the influence function of the NP-**O** estimator (see (4)). This implies that the NP-**O** estimator of $\chi_a(P;\mathcal{G})$ attains the semiparametric variance bound for $\chi_a(P;\mathcal{G})$ under $\mathcal{M}(\mathcal{G})$. For such DAG configurations there is no loss of efficiency in ignoring the observations on the variables $\mathbf{V}\setminus[\mathbf{O}\cup\{Y,A\}]$. Furthermore no loss of efficiency is incurred from ignoring the restrictions implied by the Bayesian Network $\mathcal{M}(\mathcal{G})$ on the marginal law of $(\mathbf{O},A,Y)$.

In light of the results of the preceding section, throughout we will assume that $\text{irrel}\,(A, Y, \mathcal{G}) = \emptyset$ and therefore that we can topologically sort $\mathbf{V}$ as $(W_1, \ldots, W_J, A, M_1, \ldots, M_K, Y)$, where the vertices in $\mathbf{M} = \{M_1, \ldots, M_K\}$ intersect at least one causal path between A and Y and $\mathbf{W} = \{W_1, \ldots, W_J\}$ are non-descendants of A. The set $\mathbf{O}\,(A, Y, \mathcal{G}) \equiv \mathbf{O} \equiv (O_1, \ldots, O_T)$, where $(O_1, \ldots, O_T)$ is sorted topologically, is included in $\mathbf{W}$. Throughout $T = 0$ if $\mathbf{O}\,(A, Y, \mathcal{G}) = \emptyset$, $K = 0$ if $\mathbf{M} = \emptyset$ and $J = 0$ if $\mathbf{W} = \emptyset$.

The following theorem provides necessary and sufficient conditions on the DAG $\mathcal{G}$ for (20) to hold.

Theorem 19 *If $J \in \{0, 1\}$ and $K = 0$, assertion (20) always holds. Furthermore, if $J > 1$ or $K > 0$, assertion (20) holds if and only if*

1. $\mathbf{O} \setminus \{O_T\} \subset \text{pa}_{\mathcal{G}}(O_T)$,
2. $\text{pa}_{\mathcal{G}}\,(W_{j+1}) \subset \text{pa}_{\mathcal{G}}\,(W_j) \cup \{W_j\}$ *for* $j = 1, \ldots, J-1$,
3. $\{A\} \cup \mathbf{O}_{min} \subset \text{pa}_{\mathcal{G}}(Y)$, *and*
4. $\text{pa}_{\mathcal{G}}\,(M_k) \subset \text{pa}_{\mathcal{G}}\,(M_{k-1}) \cup \{M_{k-1}\}$ *for* $k = 2, \ldots, K+1$,

where $M_{K+1} \equiv Y$. Condition 2) is nonexistent if $J \in \{0, 1\}$ and condition 4) is nonexistent if $K = 0$.

In Section 6.4 we provide an algorithm (Algorithm 2) that checks whether the necessary and sufficient conditions of Theorem 19 hold. The algorithm return false if and only if at least one of the conditions doesn't hold. As we explain in Section 6.4, when the algorithm returns false it also returns a possibly simplified formula for the efficient influence function.

Remark 20 *We emphasize that there exist DAGs such that $\chi^1_{P,a,eff}$ depends only on $\mathbf{O}, A$ and Y but the NP-$\mathbf{O}$ estimator is inefficient. One instance of this situation is given by the DAG in Figure 6. In particular, failure of one ore more of the conditions 1)-4) in Theorem 19 does not necessarily imply that the efficient estimator depends on variables other than $\mathbf{O}, A$ and Y.*

We now state a number of conditions that are implied by conditions 1)-4) of Theorem 19 and are therefore necessary for the NP-$\mathbf{O}$ estimator to be efficient. These conditions are straightforward to check. We will use them in the next examples to argue that under some DAGs the NP-$\mathbf{O}$ estimator is not efficient.

1. In Lemma 30 of the Appendix we show that for $J > 1$, whenever conditions 1) and 2) of Theorem 19 hold then

$$W_j \in \text{pa}_{\mathcal{G}}(W_{j+1}) \text{ for all } j \in \{1, \ldots, J-1\}. \tag{21}$$

   Consequently, if $J > 1$ then (21) is necessary for the NP-$\mathbf{O}$ estimator to be efficient

2. In Lemma 31 we show that if $K \geq 1$, whenever condition 4) of Theorem 19 holds for $k \in \{2, \ldots, K+1\}$ then

$$A \in \text{pa}_{\mathcal{G}}(M_1) \text{ and } M_k \in \text{pa}_{\mathcal{G}}(M_{k+1}) \text{ for } k \in \{2, \ldots, K+1\}. \tag{22}$$

   Consequently, if $K \geq 1$ then (22) is necessary for the NP-$\mathbf{O}$ estimator to be efficient.

3. The preceding two remarks imply that if either the variables in $\mathbf{W}$ or the variables in $\mathbf{M}$ can be topologically sorted in more than one way, then the NP-$\mathbf{O}$ estimator is inefficient. Thus the existence of a unique topological order of $\mathbf{W}$ and of $\mathbf{M}$ is a necessary condition for the NP-$\mathbf{O}$ estimator to be efficient.

4. Because condition 3) of Theorem 19 requires that $\mathbf{O}_{min} \subset \text{pa}_{\mathcal{G}}(Y)$ then condition 4) of the same theorem implies that

$$\mathbf{O}_{min} \subset \text{pa}_{\mathcal{G}}(M_k), \text{ for } k = 1, \ldots, K+1 \tag{23}$$

is a necessary condition for the NP-$\mathbf{O}$ estimator to be efficient.

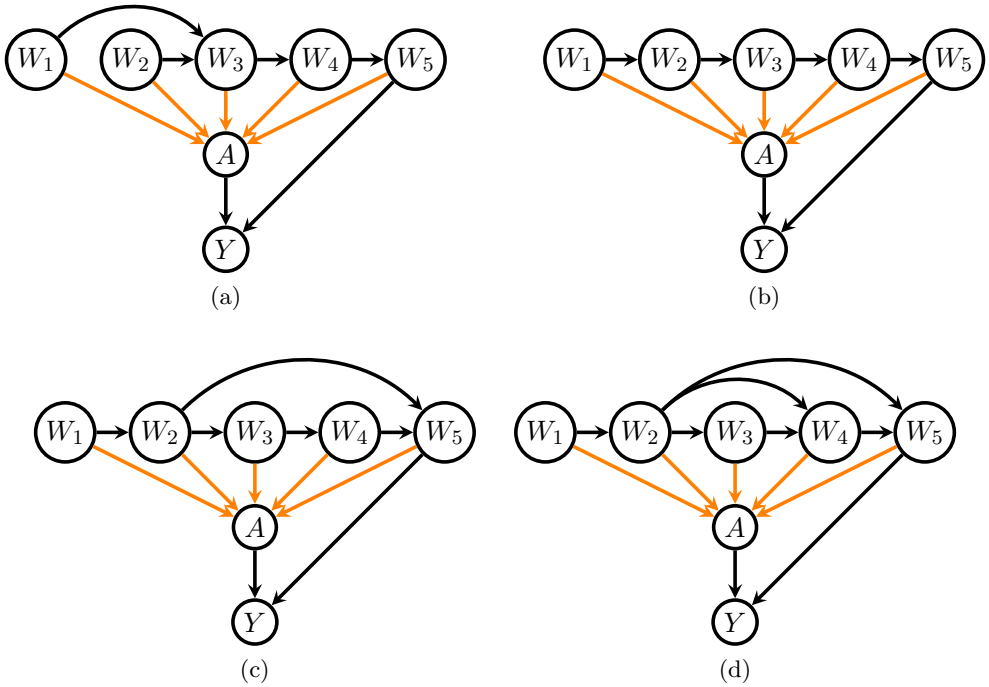

Figure 10: In graphs (a) and (c) the NP-$\mathbf{O}$ estimator is inefficient and in graphs (b) and (d) the NP-$\mathbf{O}$ estimator is efficient.

Example 8 *Consider the graphs in Figure 10. In all four graphs,* $\mathbf{M} = \emptyset$ *and* $\mathbf{O}$ *is comprised of the single node* W_5*. Thus, to check whether or not the NP-*$\mathbf{O}$ *estimator is efficient, we only need to check conditions 2) and 3) of Theorem 19. In the graph in Figure 10 (a), condition 2) fails for* $j = 2$*, because* $\text{pa}_{\mathcal{G}}(W_3) = \{W_1, W_2\}$ *which is not a subset of* $\text{pa}_{\mathcal{G}}(W_2) \cup \{W_2\} = \{W_2\}$*. Thus, the NP-*$\mathbf{O}$ *estimator is inefficient. Notice that in this graph* $\mathbf{W}$ *can be sorted topologically in two ways,* $(W_1, W_2, W_3, W_4, W_5)$ *and* $(W_2, W_1, W_3, W_4, W_5)$*. In graph 10 (b), conditions 2) and 4) hold, thus the NP-*$\mathbf{O}$ *estimator is efficient. In graph 10 (c), condition 2) fails for* $j = 4$*, because* $\text{pa}_{\mathcal{G}}(W_5) = \{W_2, W_4\}$ *which is not a subset of* $\{W_4\} \cup \text{pa}_{\mathcal{G}}(W_4)$*. Notice that in this graph* (21) *holds, thus illustrating the fact that this condition is necessary but not sufficient for the NP-*$\mathbf{O}$ *estimator to be efficient. In graph*

*10 (d), conditions 2) and 3) hold, and thus the NP-***O** *estimator is efficient. A comparison of Figures 10 (c) and 10 (d) highlights a key property that graphs for which the NP-***O** *is efficient must satisfy. Specifically, if a node W_j is a parent of a node W_{j+l} for $l > 1$ then it must also be a parent of all the nodes W_{j+h} for $1 \leq h < l$. In the graph in Figure 10 (c), W_2 is a parent of W_5 but not a parent of W_4, and thus this condition fails. All preceding assertions are valid even if any of the orange arrows from the nodes in* $\mathbf{W}$ *to A are absent.*

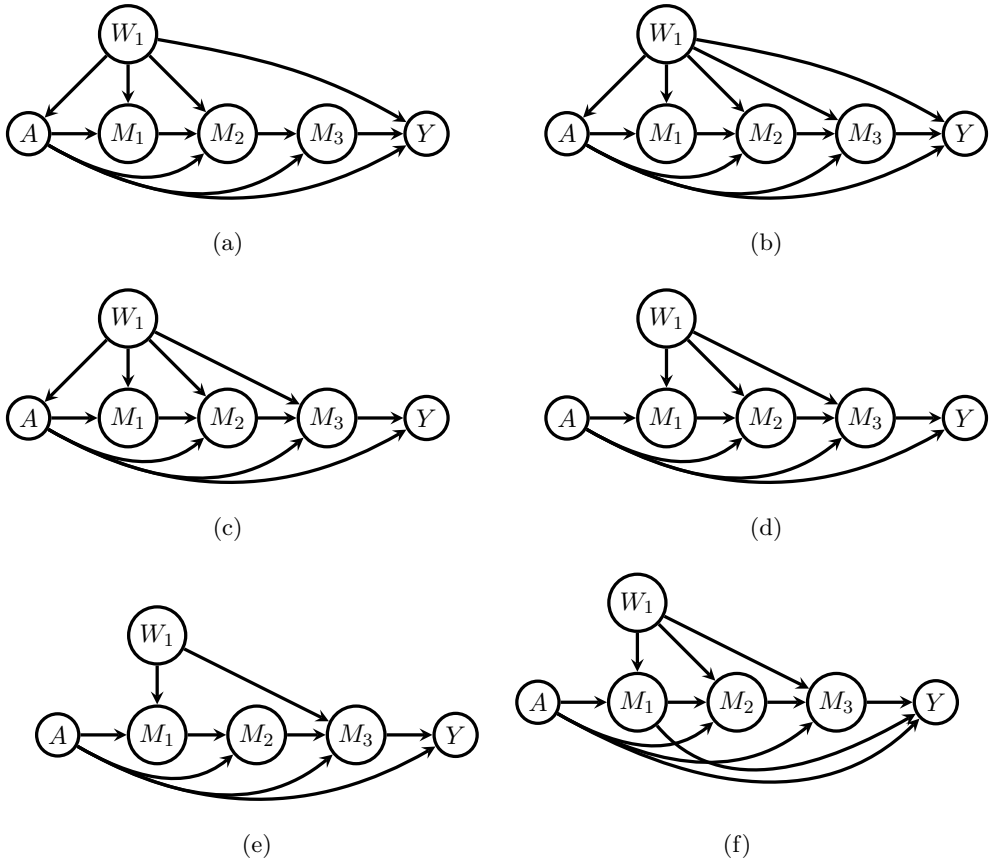

Figure 11: In graphs (a), (c), (e) and (f) the NP-**O** estimator is inefficient and in graphs (b), (d) the NP-**O** estimator is efficient.

Example 9 *Consider the graphs in Figure 11. In all four graphs,* $\mathbf{W} = \mathbf{O}$ *is comprised of the single node W_1. Furthermore in graphs (a), (b) and (c),* $\mathbf{O} = \mathbf{O}_{min}$ *and in graphs (d), (e) and (f)* $\mathbf{O}_{min} = \emptyset$*. Also,* $\mathbf{M} = \{M_1, M_2, M_3\}$*. Since $J = 1$, to check whether or not the NP-***O** *estimator is efficient, we only need to check conditions 3) and 4) of Theorem 19.*

*In graph (a), the NP-***O** *estimator is inefficient because $W_1 \notin \mathrm{pa}_{\mathcal{G}}(M_3)$, thus invalidating the necessary condition* (23)*. In graph (b), conditions 3) and 4) of Theorem 19 hold and thus the NP-***O** *estimator is efficient. In graph (c), condition 3) fails because* $\mathbf{O}_{min} = \{W_1\}$*, but $W_1 \notin \mathrm{pa}_{\mathcal{G}}(Y)$. Thus, the NP-***O** *estimator is inefficient. In graph (d), condition 3) holds because* $\mathbf{O}_{min} = \emptyset$ *and $A \in \mathrm{pa}_{\mathcal{G}}(Y)$. Furthermore, condition 4) holds. Thus, the*

*NP-***O** *estimator is efficient. In graph (e), condition 3) holds for the same reasons as in graph d). However, condition 4) fails because* $\text{pa}_{\mathcal{G}}(M_3) = \{W_1, M_2\}$ *is not included in* $\{M_2\} \cup \text{pa}_{\mathcal{G}}(M_2) = \{M_1, M_2\}$*. Thus, the NP-***O** *estimator is inefficient. In graph (f) the NP-***O** *estimator is inefficient because condition 4) fails since* $\text{pa}_{\mathcal{G}}(Y) = \{A, M_1, M_3\}$ *is not included in* $\{M_3\} \cup \text{pa}_{\mathcal{G}}(M_3) = \{W_1, M_2, M_3\}$.

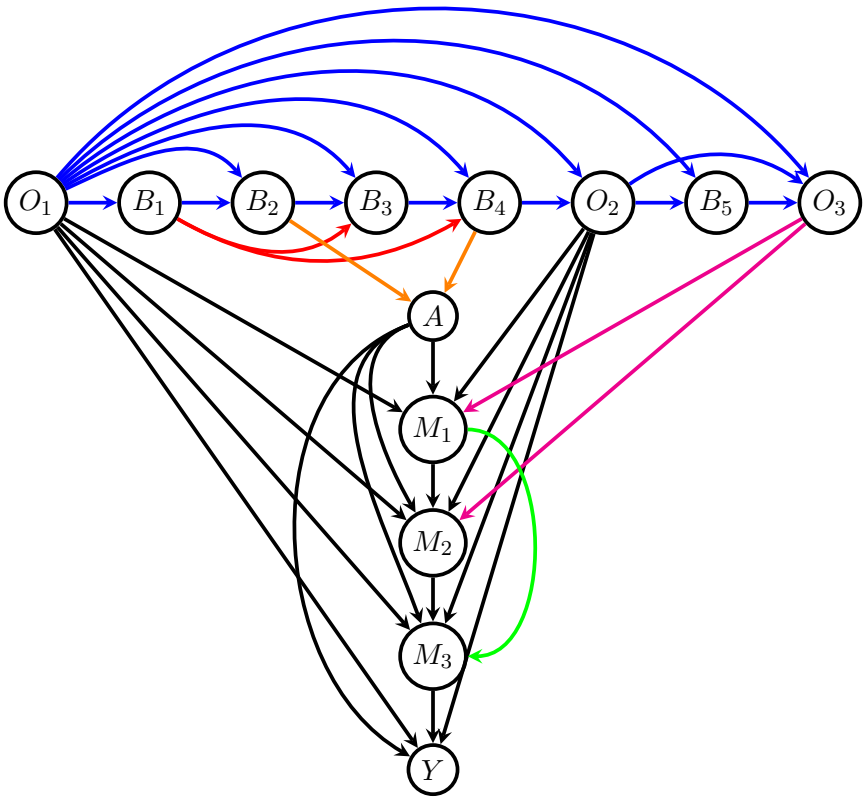

Figure 12: A DAG where the NP-**O** estimator is efficient.

Example 10 *Consider the DAG in Figure 12. In this DAG,* $J = 8$, $T = 3$,

$$\mathbf{W} = (W_1, W_2, W_3, W_4, W_5, W_6, W_7, W_8) = (O_1, B_1, B_2, B_3, B_4, O_2, B_5, O_3)\,,$$

$\mathbf{O} = \{O_1, O_2, O_3\}$ *and* $\mathbf{O}_{min} = \{O_1, O_2\}$*. It can be checked that conditions 1)-4) of Theorem 19 hold for this graph and thus the NP-***O** *estimator is efficient. All blue arrows in the graph are needed for condition 1) and 2) to hold. That is, if any blue arrow were absent from the graph, the NP-***O** *estimator would be inefficient. If both red arrows were absent or if only the red arrow from* B_1 *to* B_3 *was present, then condition 2) would hold. Notice that condition 1) does not impose any restriction to the presence or absence of the red arrows. However, if the arrow from* B_1 *to* B_4 *were present but the arrow from* B_1 *to* B_3 *were absent, then*

condition 2) would fail. None of the conditions 1)-4) impose restrictions on the existence of edges connecting nodes in $\mathbf{W} \setminus \mathbf{O}$ *to* A*. Thus, the conclusions of our discussion would remain valid regardless of whether or not any of the orange edges are present. All black arrows are needed for conditions 3) and 4) to hold. The requirements on the presence or absence of the purple arrows connecting* O_3 *to* M_1 *and* M_2 *are similar to the requirements on the red arrows. Specifically, condition 4) would remain valid even if the edge connecting* O_3 *to* M_1 *or both the edges connecting* O_3 *with* M_1 *and with* M_2 *were absent. However, if the edge connecting* O_3 *with* M_2 *is present then the edge connecting* O_3 *with* M_1 *must be present for condition 4) to hold. Finally, the conditions of Theorem 19 would remain valid even if the green edge were missing.*

6.3 A Connection Between Identification and Efficient NP-O Estimation

Theorem 19 has the following interesting corollary.

Theorem 21 *Let* $\mathcal{G}$ *be a DAG with vertex set* $\mathbf{V}$*, let* A *and* Y *be two distinct vertices in* $\mathbf{V}$ *with* A *corresponding to a point intervention. Let* $\mathbf{O} = \mathbf{O}(A, Y, \mathcal{G})$ *and let* $\mathbf{O}_{\min} \subset \mathbf{O}$ *be the subset of* $\mathbf{O}$ *with the smallest number of vertices such that* $A \perp\!\!\!\perp_{\mathcal{G}} [\mathbf{O} \backslash \mathbf{O}_{\min}] \,|\mathbf{O}_{\min}$*. Suppose that* $\mathbf{O}_{min} \neq \emptyset$*. Let* $\mathbf{M} = \mathrm{cn}(A, Y, \mathcal{G}) \setminus \{Y\}$*. If there exists an identifying formula for* $\chi_a(P; \mathcal{G})$ *that depends only on* A, Y *and the mediators* $\mathbf{M}$ *then the NP-*$\mathbf{O}$ *estimator of* $\chi_a(P; \mathcal{G})$ *is not globally efficient under the Bayesian Network* $\mathcal{M}(\mathcal{G})$*.*

Proof We prove the result by contradiction. By Lemma 16, without loss of generality, we can assume that $\mathrm{irrel}(A, Y, \mathbf{G}) = \emptyset$. Suppose that the NP-$\mathbf{O}$ estimator of $\chi_a(P; \mathcal{G})$ is globally efficient under $\mathcal{M}(\mathcal{G})$. Then Theorem 19 implies that every vertex $\mathbf{O}_{min}$ must be a parent of Y and of every vertex in $\mathbf{M}$. Furthermore, by (22), A must be a parent of M_1 and each M_k must be a parent of M_{k+1}. Let $\mathcal{G}\left[\{A, Y\} \cup \mathbf{M}\right]$ be the latent projection of $\mathcal{G}$ (Evans and Richardson, 2014) onto the vertex set $\{A, Y\} \cup \mathbf{M}$. Because $\mathbf{O}_{min}$ is not empty, then, in $\mathcal{G}\left[\{A, Y\} \cup \mathbf{M}\right]$, the nodes $A, M_1, M_2, \ldots, M_K, Y$ are all in the same district (Shpitser et al., 2014). The completeness of the ID algorithm, see Tian and Pearl (2002) and Shpitser and Pearl (2008), implies that $\chi_a(P; \mathcal{G})$ is not identified when only $\{A, Y\} \cup \mathbf{M}$ are observed, thus arriving at a contradiction. ■

Interestingly, Theorem 21 implies that in the front-door DAG in Figure 8, the NP-$\mathbf{O}$ estimator is not efficient.

6.4 A Sound Algorithm to Check for Variables That Are Not Needed For Semiparametric Efficient Estimation

In the Section 6.2 we showed that for certain DAG configurations, the efficient influence function $\chi^1_{P,a,eff}(\mathbf{V}; \mathcal{G})$ depends on $\mathbf{V}$ only through $\mathbf{O}, A$ and Y. In this section we argue that when $\chi^1_{P,a,eff}(\mathbf{V}; \mathcal{G})$ depends on variables in $\mathbf{V} \backslash [\mathbf{O} \cup \{A, Y\}]$ it may nevertheless depend on a strict subset of $\mathbf{V}$. As such, from the perspective of planning a study, it is useful to learn which variables are irrelevant for efficient estimation since such variables need not be measured.

As an example, consider the DAG in Figure 10 (c). Applying formula (19) from Theorem 18, the d-separations in the DAG imply that

$$\begin{aligned}\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G}) = &\frac{I_a(A)}{\pi_a(W_5;P)}Y - \frac{I_a(A)}{\pi_a(W_5;P)}b_a\left(W_5;P\right)\\ &+ b_a(W_5;P) - \chi_a(P;\mathcal{G}) + E_P\left[b_a(W_5;P) \mid W_3, W_4\right] - E_P\left[b_a(W_5;P) \mid W_2, W_4\right]\\ &+ E_P\left[b_a(W_5;P) \mid W_2, W_3\right] - E_P\left[b_a(W_5;P) \mid W_3\right].\end{aligned}$$

Check Example 11 in the Appendix for the details involved in the preceding calculation. Notice that W_1 does not enter into the formula for $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$. Thus W_1 can be ignored for efficient estimation of $\chi_a(P;\mathcal{G})$.

Algorithm 2 provides a sound check for possible simplifications of $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$. Such simplifications may imply that some variables do not appear in the formula for the efficient influence function, as is the case in the example of the DAG of Figure 10 (c) we just discussed. Whether or not the algorithm returns the smallest subset of $\mathbf{V}$ on which $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$ depends on remains an open problem. Algorithm 2 is also sound and complete for the query of whether $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$ simplifies to $\psi_{P,a}[\mathbf{O};\mathcal{G}]$. Proofs of all these assertions together with a heuristic explanation of the rationale for each step in the algorithm, are given in the Appendix.

Algorithm 2: Sound and complete check of $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G}) = \psi_{P,a}(\mathbf{O};\mathcal{G})$

input : DAG $\mathcal{G}$ with vertex set $\mathbf{V}$ and two distinct nodes $A, Y \in \mathbf{V}$ such that $A \in \text{an}_{\mathcal{G}}(Y)$

output: An answer to the inquiry of whether $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G}) = \psi_{P,a}(\mathbf{O};\mathcal{G})$ for all $P \in \mathcal{M}(\mathcal{G})$ and a, possibly, simplified formula for $\chi^1_{P,a,eff}$ if the answer to the inquiry is negative.

procedure `checkEfficient`$(A, Y, \mathcal{G})$
- $\mathcal{G}$=`prune`$(A, Y, \mathcal{G})$
- $(\mathbf{W}, A, \mathbf{M}, Y) = (W_1, \ldots, W_J, A, M_1, \ldots, M_K, Y)$ =`topological_sort`$(\mathbf{V}, \mathcal{G})$
- /* $J = 0$ `if` $\mathbf{W} = \emptyset$ `and` $K = 0$ `if` $\mathbf{M} = \emptyset$ */
- $M_{K+1} = Y$
- $\mathbf{O} = \mathbf{O}(A, Y, \mathcal{G})$
- $O_1, \ldots, O_T$ =`topological_sort`$(\mathbf{O})$
- `efficient_nondesc`=`False`
- `efficient_desc`=`False`
- **if** $\mathbf{O} \setminus \{O_T\} \subset \text{pa}_{\mathcal{G}}(O_T)$ *and* $J > 1$ **then**
  - $j = J - 1$
  - **while** $\text{pa}_{\mathcal{G}}(W_{j+1}) \setminus \{W_j\} \subset \text{pa}_{\mathcal{G}}(W_j)$ *and* $j \geq 2$ **do** $j = j - 1$
  - **if** $j \geq 2$ **then**
    - `offenders_nondesc`=$\{j\}\cup$`get_offenders_nondesc`$(\mathcal{G}, \mathbf{W}, \mathbf{O}, j-1)$
    - $\chi^{1,non-desc}_{P,a,eff} = b_a(\mathbf{O};P) - \chi_a(P;\mathcal{G}) + \sum_{h\in\texttt{offenders\_nondesc}} \left\{E_P\left[b_a(\mathbf{O};P) \,|\, \text{pa}_{\mathcal{G}}(W_h), W_h\right] - E_P\left[b_a(\mathbf{O};P) \,|\, \text{pa}_{\mathcal{G}}(W_{h+1})\right]\right\}$
  - **else**
    - $\chi^{1,non-desc}_{P,a,eff} = b_a(\mathbf{O};P) - \chi_a(P;\mathcal{G})$
    - `efficient_nondesc`=`True`
- **else if** $J > 1$ **then**
  - `offenders_nondesc`=`get_offenders_nondesc`$(\mathcal{G}, \mathbf{W}, \mathbf{O}, J-1)$
  - $\chi^{1,non-desc}_{P,a,eff} = E_P\left[b_a(\mathbf{O};P) \mid W_J, \text{pa}_{\mathcal{G}}(W_J)\right] - \chi_a(P;\mathcal{G}) + \sum_{h\in\texttt{offenders\_nondesc}} \left\{E_P\left[b_a(\mathbf{O};P) \,|\, \text{pa}_{\mathcal{G}}(W_h), W_h\right] - E_P\left[b_a(\mathbf{O};P) \,|\, \text{pa}_{\mathcal{G}}(W_{h+1})\right]\right\}$
- **else if** $J = 1$ **then**
  - $\chi^{1,non-desc}_{P,a,eff} = b_a(\mathbf{O};P) - \chi_a(P;\mathcal{G})$
  - `efficient_nondesc`=`True`
- **else if** $J = 0$ **then**
  - $\chi^{1,non-desc}_{P,a,eff} = 0$
  - `efficient_nondesc`=`True`
- **if** $A \cup \mathbf{O}_{min} \subset \text{pa}_{\mathcal{G}}(Y)$ *and* $K \geq 1$ **then**
  - $\chi^{1,desc}_{P,a,eff} = I_a(A) Y \pi_a^{-1}(\mathbf{O}_{min};P)$
  - $k = K + 1$
  - **while** $k \geq 2$ && $\text{pa}_{\mathcal{G}}(M_k) \subset \text{pa}_{\mathcal{G}}(M_{k-1}) \cup \{M_{k-1}\}$ **do** $k = k - 1$
  - **if** $k \geq 2$ **then**
    - `offenders_desc`=$\{k\}\cup$`get_offenders_desc`$(\mathcal{G}, \mathbf{M}, \mathbf{O}, \mathbf{O}_{min}, k-1)$
    - $\chi^{1,desc}_{P,a,eff} = \chi^{1,desc}_{P,a,eff} + \sum_{h\in\texttt{offenders\_desc}}\{E_P\left[T_{P,a,\mathcal{G}} \mid \text{pa}_{\mathcal{G}}(M_{h-1}), M_{h-1}\right] - E_P\left[T_{P,a,\mathcal{G}} \mid \text{pa}_{\mathcal{G}}(M_h)\right]\}$
    - **if** $\{A\} \cup \mathbf{O} = \text{pa}_{\mathcal{G}}(M_1)$ **then**
      - $\chi^{1,desc}_{P,a,eff} = \chi^{1,desc}_{P,a,eff} - I_a(A) b_a(\mathbf{O};P) \pi_a^{-1}(\mathbf{O}_{min};P)$
    - **else**
      - $\chi^{1,desc}_{P,a,eff} = \chi^{1,desc}_{P,a,eff} - E_P\left[T_{P,a,\mathcal{G}} \mid \text{pa}_{\mathcal{G}}(M_1)\right]$
  - **else**
    - $\chi^{1,desc}_{P,a,eff} = \chi^{1,desc}_{P,a,eff} - I_a(A) b_a(\mathbf{O};P) \pi_a^{-1}(\mathbf{O}_{min};P)$
    - `efficient_desc=True`
- **else if** $K \geq 1$ **then**
  - `offenders_desc`=$\{K+1\}\cup$`get_offenders_desc`$(\mathcal{G}, \mathbf{M}, \mathbf{O}, \mathbf{O}_{min}, K)$
  - $\chi^{1,desc}_{P,a,eff} = \sum_{h\in\texttt{offenders\_desc}}\{E_P\left[T_{P,a,\mathcal{G}} \mid \text{pa}_{\mathcal{G}}(M_{h-1}), M_{h-1}\right] - E_P\left[T_{P,a,\mathcal{G}} \mid \text{pa}_{\mathcal{G}}(M_h)\right]\}$
  - **if** $\{A\} \cup \mathbf{O} = \text{pa}_{\mathcal{G}}(M_1)$ **then**
    - $\chi^{1,desc}_{P,a,eff} = \chi^{1,desc}_{P,a,eff} - I_a(A) b_a(\mathbf{O};P) \pi_a^{-1}(\mathbf{O}_{min};P)$
  - **else**
    - $\chi^{1,desc}_{P,a,eff} = \chi^{1,desc}_{P,a,eff} - E_P\left[T_{P,a,\mathcal{G}} \mid \text{pa}_{\mathcal{G}}(M_1)\right]$
- **else**
  - $\chi^{1,desc}_{P,a,eff} = I_a(A) \pi_a^{-1}(\mathbf{O}_{min};P)(Y - b_a(\mathbf{O};P))$
  - `efficient_desc=True`
- $\chi^1_{P,a,eff} = \chi^{1,non-desc}_{P,a,eff} + \chi^{1,desc}_{P,a,eff}$
- `efficient` = `efficient_desc` & `efficient_nondesc`
- **return** `efficient`, $\chi^1_{P,a,eff}$

Algorithm 3: Subroutine to find all mediator nodes that don't satisfy at least one of (66)**,** (69) **or** (70)**.**

input : DAG $\mathcal{G}$, mediator nodes $\mathbf{M}$, optimal adjustment set $\mathbf{O}$, optimal minimal adjustment set $\mathbf{O}_{min}$ and integer init.
output: The set of nodes in $\mathbf{M}$ that don't satisfy at least one of (66), (69) or (70)
procedure `get_offenders_desc`$(\mathcal{G}, \mathbf{M}, \mathbf{O}, \mathbf{O}_{min}, \text{init})$
  `offender_desc`$= \emptyset$
  for $i = \text{init}, \ldots, 2$ **do**
    if $\{A\} \cup \mathbf{O}_{min} \not\subset \text{pa}_{\mathcal{G}}(M_1)$ *or* $\text{pa}_{\mathcal{G}}(M_i) \not\subset \text{pa}_{\mathcal{G}}(M_{i-1}) \cup \{M_{i-1}\}$ *or* $Y \not\perp\!\!\!\perp_{\mathcal{G}} \text{pa}_{\mathcal{G}}(M_{i-1}) \cup \{M_{i-1} \setminus \text{pa}_{\mathcal{G}}(M_i) \mid \text{pa}_{\mathcal{G}}(M_i)$ **then**
      `offender_desc`$=$`offender_desc`$\cup\{i\}$
  return `offender_desc`

Algorithm 4: Subroutine to find all non-mediator nodes that don't satisfy (60).

input : DAG $\mathcal{G}$, non-mediator nodes $\mathbf{W}$, optimal adjustment set $\mathbf{O}$ and integer init.
output: The set of nodes in $\mathbf{W}$ that don't satisfy (60)
procedure `get_offenders_nondesc`$(\mathcal{G}, \mathbf{W}, \mathbf{O}, \text{init})$
  `offender_nondesc`$= \emptyset$
  for $i = \text{init}, \ldots, 1$ **do**
    $\mathbf{I}_i \equiv \left[\text{pa}_{\mathcal{G}}(W_i) \cup \{W_i\}\right] \cap \text{pa}_{\mathcal{G}}(W_{i+1})$
    if $\mathbf{O} \setminus \mathbf{I}_i \not\perp\!\!\!\perp_{\mathcal{G}} \left[\text{pa}_{\mathcal{G}}(W_i) \cup \{W_i\}\right] \triangle \text{pa}_{\mathcal{G}}(W_{i+1}) \mid \mathbf{I}_i$ **then**
      `offender_nondesc`$=$`offender_nondesc`$\cup\{i\}$
  return `offender_nondesc`

7. Discussion

The results in this paper raise a number of open problems, several of which we are currently investigating.

1. The derivation of a graphical criterion to characterize the class of all time dependent adjustment sets, like adjustment sets in row 1 and 8 in Example 4, that dominate the rest even if they don't dominate each other.

2. The characterization of DAGs under which an optimal time dependent adjustment set exists for joint interventions.

3. The characterization of the subset of DAGs such that an optimal time dependent adjustment set exists for joint interventions, and for which the optimal time dependent adjustment set is time independent.

4. The characterization of DAGs such that among the adjustment sets of minimal size, there exists an optimal one.

5. For DAGs for which an optimal time dependent adjustment set exists, the derivation of a sound and complete algorithm to answer the inquiry of whether the non-parametric optimally adjusted estimator is globally efficient under the Bayesian Network.

6. For DAGs with latent variables such that observable adjustment sets exist, the characterization of the subset of DAGs for which an optimal adjustment set exists among the observable adjustment sets.

7. For DAGs with latent variables, the derivation of a general expression for the efficient influence function of $\chi_a(P;\mathcal{G})$ and a non-parametric globally efficient estimator.

8. For DAGs with latent variables for which an optimal observable adjustment set exists, the derivation of a sound and complete algorithm to answer the inquiry of whether the non-parametric optimally adjusted estimator is globally efficient under the marginal of Bayesian Network for the observable variables.

Acknowledgments

The authors wish to thank three anonymous referees and the action editor for helpful and insightful comments which helped to greatly improve the presentation of the paper. Ezequiel Smucler was partially supported by Grant 20020170100330BA from Universidad de Buenos Aires and PICT-201-0377 from ANPYCT, Argentina.

Appendix A. Main proofs

A.1 Proofs of results in Section 5

Proof [Proof of Lemma 4] We first show that $(\mathbf{G}, \mathbf{B})$ is an adjustment set. Since $\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{G} \mid \mathbf{B}$ we have

$$\pi_{\mathbf{a}}(\mathbf{G}, \mathbf{B}; P) = \pi_{\mathbf{a}}(\mathbf{B}; P). \tag{24}$$

Then, for all $P \in \mathcal{M}(\mathcal{G})$, $E_P\left[\frac{I_{\mathbf{a}}(\mathbf{A})Y}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)}\right] = E_P\left[\frac{I_{\mathbf{a}}(\mathbf{a})Y}{\pi_{\mathbf{a}}(\mathbf{B};P)}\right] = \chi_{\mathbf{a}}(P;\mathcal{G})$ where the last equality holds because $\mathbf{B}$ is by assumption an adjustment set. This shows that $(\mathbf{G}, \mathbf{B})$ is an adjustment set. Now,

$$\begin{aligned}
\psi_{P,\mathbf{a}}(\mathbf{B};\mathcal{G}) &= \frac{I_{\mathbf{a}}(\mathbf{A})Y}{\pi_{\mathbf{a}}(\mathbf{B};P)} - \left[\frac{I_{\mathbf{a}}(\mathbf{A})}{\pi_{\mathbf{a}}(\mathbf{B};P)} - 1\right] b_{\mathbf{a}}(\mathbf{B};P) - \chi_{\mathbf{a}}(P;\mathcal{G}) \\
&= \frac{I_{\mathbf{a}}(\mathbf{A})Y}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)} - \left[\frac{I_{\mathbf{a}}(\mathbf{A})}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)} - 1\right] b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P) - \chi_{\mathbf{a}}(P;\mathcal{G}) \\
&+ \left[\frac{I_{\mathbf{a}}(\mathbf{A})}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)} - 1\right] \{b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P) - b_{\mathbf{a}}(\mathbf{B};P)\} \\
&= \psi_{P,\mathbf{a}}(\mathbf{G},\mathbf{B};\mathcal{G}) + \left[\frac{I_{\mathbf{a}}(\mathbf{A})}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)} - 1\right] [b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P) - b_{\mathbf{a}}(\mathbf{B};P)]
\end{aligned}$$

where the second equality follows from (24) . Next, noting that

$$E_P\{\psi_{P,\mathbf{a}}[\mathbf{G},\mathbf{B};\mathcal{G}]\, g(\mathbf{A},\mathbf{G},\mathbf{B})\} = 0 \text{ for any } g \text{ such that } E_P[g(\mathbf{A},\mathbf{G},\mathbf{B})|\mathbf{G},\mathbf{B}] = 0 \tag{25}$$

and that $E_P\left\{\left[\frac{I_{\mathbf{a}}(\mathbf{A})}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)} - 1\right][b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P) - b_{\mathbf{a}}(\mathbf{B};P)]\middle|\, \mathbf{G},\mathbf{B}\right\} = 0$ we conclude that

$$\begin{aligned}
&\sigma^2_{\mathbf{a},\mathbf{B}}(P) \equiv var_P[\psi_{P,\mathbf{a}}(\mathbf{B};\mathcal{G})] \qquad (26)\\
&= var_P[\psi_{P,\mathbf{a}}(\mathbf{G},\mathbf{B};\mathcal{G})] + var_P\left[\left\{\frac{I_{\mathbf{a}}(\mathbf{A})}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)} - 1\right\}\{b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P) - b_{\mathbf{a}}(\mathbf{B};P)\}\right] \\
&\equiv \sigma^2_{\mathbf{a},\mathbf{G},\mathbf{B}}(P) + E_P\left\{E\left[[b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P) - b_{\mathbf{a}}(\mathbf{B};P)]^2\middle|\,\mathbf{B}\right] var_P\left[\frac{I_{\mathbf{a}}(\mathbf{A})}{\pi_{\mathbf{a}}(\mathbf{B};P)} - 1 \mid \mathbf{B}\right]\right\} \\
&= \sigma^2_{\mathbf{a},\mathbf{G},\mathbf{B}}(P) + E_P\left\{var_P(b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P) \mid \mathbf{B})\left[\frac{1}{\pi_{\mathbf{a}}(\mathbf{B};P)} - 1\right]\right\}
\end{aligned}$$

where the last equality holds because $b_{\mathbf{a}}(\mathbf{B};P) = E_P[b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P) \mid \mathbf{A}=\mathbf{a},\mathbf{B}] = E_P[b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P) \mid \mathbf{B}]$, the first equality is by the definitions of $b_{\mathbf{a}}(\mathbf{B},P)$ and $b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)$ and the second is true because $\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{G} \mid \mathbf{B}$. Next, recall that $\mathbf{c} \equiv (c_{\mathbf{a}})_{\mathbf{a}\in\mathbf{A}}$, $\mathbf{Q} \equiv [Q_a]_{\mathbf{a}\in\mathbf{A}}$ where $Q_{\mathbf{a}} \equiv \left\{\frac{I_{\mathbf{a}}(\mathbf{A})}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)} - 1\right\}\{b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P) - b_{\mathbf{a}}(\mathbf{B};P)\}$. For any $\mathbf{Z}$, define $\psi_P(\mathbf{Z};\mathcal{G}) \equiv (\psi_{P,\mathbf{a}}(\mathbf{Z};\mathcal{G}))_{\mathbf{a}\in\mathbf{A}}$. Then, writing $\sum_{\mathbf{a}\in\mathbf{A}} c_{\mathbf{a}}\psi_{P,\mathbf{a}}(\mathbf{Z};\mathcal{G}) = \mathbf{c}^T\psi_P(\mathbf{Z};\mathcal{G})$ and noticing that $E_P[\mathbf{Q}|\mathbf{G},\mathbf{B}] = \mathbf{0}$, it follows from (25) that $\sigma^2_{\Delta,\mathbf{B}}(P) = var_P\left[\mathbf{c}^T\psi_P(\mathbf{B};\mathcal{G})\right] = var_P\left[\mathbf{c}^T\psi_P(\mathbf{G},\mathbf{B};\mathcal{G})\right] + var_P\left[\mathbf{c}^T\mathbf{Q}\right] = \sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P) + \mathbf{c}^T var_P(\mathbf{Q})\,\mathbf{c}$. The expression for $var_P(Q_{\mathbf{a}})$ was derived in (26). On the other

hand if $\mathbf{a} \neq \mathbf{a}\prime$

$$
\begin{aligned}
&cov_P\left(Q_{\mathbf{a}}, Q_{\mathbf{a}'}\right) = \\
&E_P\left[\left\{\frac{I_{\mathbf{a}}(\mathbf{A})}{\pi_{\mathbf{a}}(\mathbf{B};P)} - 1\right\}\left\{\frac{I_{\mathbf{a}'}(\mathbf{A})}{\pi_{\mathbf{a}'}(\mathbf{B};P)} - 1\right\} cov_P\left[b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P), b_{\mathbf{a}'}(\mathbf{G},\mathbf{B};P)|\mathbf{B},\mathbf{A}\right]\right] = \\
&E_P\left[\frac{cov_P\left[I_{\mathbf{a}}(\mathbf{A}), I_{\mathbf{a}'}(\mathbf{A})|\mathbf{B}\right]}{\pi_{\mathbf{a}}(\mathbf{B};P)\pi_{\mathbf{a}'}(\mathbf{B};P)} cov_P\left[b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P), b_{\mathbf{a}'}(\mathbf{G},\mathbf{B};P)|\mathbf{B}\right]\right] = \\
&- E_P\left[cov_P\left[b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P), b_{\mathbf{a}'}(\mathbf{G},\mathbf{B};P)|\mathbf{B}\right]\right].
\end{aligned}
$$

This concludes the proof Lemma 4 ■

Proof [Proof of Lemma 5] We first show that $\mathbf{G}$ is an adjustment set. For any $P \in \mathcal{M}(\mathcal{G})$ the assumption $Y \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{A}, \mathbf{G}$ implies

$$
b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P) \equiv E_P\left(Y \mid \mathbf{A}=\mathbf{a},\mathbf{G},\mathbf{B}\right) = E_P\left(Y \mid \mathbf{A}=\mathbf{a},\mathbf{G}\right) \equiv b_{\mathbf{a}}(\mathbf{G};P) \tag{27}
$$

and consequently that $E_P\left[b_{\mathbf{a}}(\mathbf{G};P)\right] = E_P\left[b_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)\right] = \chi_{\mathbf{a}}(P;\mathcal{G})$ where the second equality follows from the assumption that $(\mathbf{G},\mathbf{B})$ is an adjustment set. This shows that $\mathbf{G}$ is an adjustment set. Now,

$$
\begin{aligned}
&E_P\left[\psi_{P,\mathbf{a}}(\mathbf{G},\mathbf{B};\mathcal{G}) \mid \mathbf{A}, Y, \mathbf{G}\right] = \\
&I_{\mathbf{a}}(\mathbf{A})\left\{Y - b_{\mathbf{a}}(\mathbf{G};P)\right\} E_P\left[\frac{1}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)}\middle| \mathbf{A}=\mathbf{a}, Y, \mathbf{G}\right] + \left\{b_{\mathbf{a}}(\mathbf{G};P) - \chi_{\mathbf{a}}(P;\mathcal{G})\right\} = \\
&I_{\mathbf{a}}(\mathbf{A})\left\{Y - b_{\mathbf{a}}(\mathbf{G};P)\right\} E_P\left[\frac{1}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)}\middle| \mathbf{A}=\mathbf{a}, \mathbf{G}\right] + \left\{b_{\mathbf{a}}(\mathbf{G};P) - \chi_{\mathbf{a}}(P;\mathcal{G})\right\} = \\
&\frac{I_{\mathbf{a}}(\mathbf{A})}{\pi_{\mathbf{a}}\left(\mathbf{G};P\right)}\left\{Y - b_{\mathbf{a}}(\mathbf{G};P)\right\} + \left\{b_{\mathbf{a}}(\mathbf{G};P) - \chi_{\mathbf{a}}(P;\mathcal{G})\right\} = \psi_{P,\mathbf{a}}(\mathbf{G};\mathcal{G}).
\end{aligned} \tag{28}
$$

where the first equality follows from (27) , the second follows from $Y \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{A}, \mathbf{G}$ and the third by invoking Lemma 27 in Section A.4. On the other hand,

$$
\begin{aligned}
&E_P\left[var_P\left[\psi_{P,\mathbf{a}}(\mathbf{G},\mathbf{B};\mathcal{G}) \mid \mathbf{A}, Y, \mathbf{G}\right]\right] = \\
&E_P\left[var_P\left\{\frac{I_{\mathbf{a}}(\mathbf{A})}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)}\left\{Y - b_{\mathbf{a}}(\mathbf{G};P)\right\} + \left\{b_{\mathbf{a}}(\mathbf{G};P) - \chi_{\mathbf{a}}(P;\mathcal{G})\right\} \mid \mathbf{A}, Y, \mathbf{G}\right\}\right] = \\
&E_P\left[I_{\mathbf{a}}(\mathbf{A})\left(Y - b_{\mathbf{a}}(\mathbf{G};P)\right)^2 var_P\left\{\frac{1}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)} \mid \mathbf{A}=\mathbf{a}, \mathbf{G}\right\}\right] = \\
&E_P\left\{\pi_{\mathbf{a}}(\mathbf{G};P) var_P(Y \mid \mathbf{A}=\mathbf{a},\mathbf{G}) var_P\left[\frac{1}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)} \mid \mathbf{A}=\mathbf{a},\mathbf{G}\right]\right\}
\end{aligned}
$$

where the first equality follows from (27) and the second follows from $Y \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{A}, \mathbf{G}$. We therefore have

$$
\begin{aligned}
&\sigma^2_{\mathbf{a},\mathbf{G},\mathbf{B}}\left(P\right) \equiv var_P\left[\psi_{P,\mathbf{a}}(\mathbf{G},\mathbf{B};\mathcal{G})\right] \\
&= var_P\left[\psi_{P,\mathbf{a}}(\mathbf{G};\mathcal{G})\right] + E_P\left\{\pi_{\mathbf{a}}(\mathbf{G};P) var_P(Y \mid \mathbf{A}=\mathbf{a},\mathbf{G}) var_P\left[\frac{1}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)} \mid \mathbf{A},\mathbf{G}\right]\right\} \\
&= \sigma^2_{\mathbf{a},\mathbf{G}}\left(P\right) + E_P\left\{\pi_{\mathbf{a}}(\mathbf{G};P) var_P(Y \mid \mathbf{A}=\mathbf{a},\mathbf{G}) var_P\left[\frac{1}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)} \mid \mathbf{A},\mathbf{G}\right]\right\}.
\end{aligned}
$$

Next,

$$\begin{aligned}\sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P) &= var_P\left[E_P\left[\mathbf{c}^T\psi_P(\mathbf{G},\mathbf{B};\mathcal{G})\,|\mathbf{A},Y,\mathbf{G}\right]\right] + E_P\left[var_P\left[\mathbf{c}^T\psi_P(\mathbf{G},\mathbf{B};\mathcal{G})\,|\mathbf{A},Y,\mathbf{G}\right]\right]\\ &= var_P\left[\mathbf{c}^T\psi_P(\mathbf{G};\mathcal{G})\right] + \mathbf{c}^T E_P\left[var_P\left[\psi_P(\mathbf{G},\mathbf{B};\mathcal{G})\,|\mathbf{A},Y,\mathbf{G}\right]\right]\mathbf{c}\\ &= \sigma^2_{\Delta,\mathbf{G}}(P) + \mathbf{c}^T E_P\left[var_P\left[\psi_P(\mathbf{G},\mathbf{B};\mathcal{G})\,|\mathbf{A},Y,\mathbf{G}\right]\right]\mathbf{c}.\end{aligned}$$

But by (27) we have $cov_P\left[\psi_{\mathbf{a},P}(\mathbf{G},\mathbf{B};\mathcal{G}),\psi_{\mathbf{a}',P}(\mathbf{G},\mathbf{B};\mathcal{G})\,|\mathbf{A},Y,\mathbf{G}\right]=0$ because $I_{\mathbf{a}}(\mathbf{A})I_{\mathbf{a}'}(\mathbf{A})=0$. Consequently

$$\mathbf{c}^T E_P\left[var_P\left[\psi_P(\mathbf{G},\mathbf{B};\mathcal{G})\,|\mathbf{A},Y,\mathbf{G}\right]\right]\mathbf{c} =$$
$$\sum_{\mathbf{a}\in\mathcal{A}} c^2_{\mathbf{a}} E_P\left\{\pi_{\mathbf{a}}(\mathbf{G};P)var_P(Y\mid\mathbf{A}=\mathbf{a},\mathbf{G})var_P\left[\frac{1}{\pi_{\mathbf{a}}(\mathbf{G},\mathbf{B};P)}\mid\mathbf{A}=\mathbf{a},\mathbf{G}\right]\right\}.$$

The formula for $\sigma^2_{ATE,\mathbf{G},\mathbf{B}}(P)-\sigma^2_{ATE,\mathbf{G}}(P)$ follows immediately by applying the formula for $\sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P)$ to $c_{a=1}=1$ and $c_{a=-1}=-1$. This concludes the proof of Lemma 5. ■

Lemma 22 *Given a DAG $\mathcal{G}$ and disjoint vertices $\mathbf{A}$ and $\mathbf{B}$ there exists a unique subset $\mathbf{C}$ of $\mathbf{B}$ such that $\mathbf{A}\perp\!\!\!\perp_{\mathcal{G}}\mathbf{B}\backslash\mathbf{C}\mid\mathbf{C}$ and such that no strict subset $\mathbf{C}'$ of $\mathbf{C}$ satisfies $\mathbf{A}\perp\!\!\!\perp_{\mathcal{G}}\mathbf{B}\backslash\mathbf{C}'\mid\mathbf{C}'$.*

Proof The result is a consequence of the fact that d-separation is a graphoid. See Geiger et al. (1990). Suppose there were two distinct minimal sets, say $\mathbf{C}_1$ and $\mathbf{C}_2$. Let $\mathbf{I}=\mathbf{C}_1\cap\mathbf{C}_2$, $\mathbf{W}_1=\mathbf{C}_1\backslash\mathbf{I}$ and $\mathbf{W}_2=\mathbf{C}_2\backslash\mathbf{I}$ and $\mathbf{R}=\mathbf{B}\backslash(\mathbf{C}_1\cup\mathbf{C}_2)$. Then $\mathbf{A}\perp\!\!\!\perp_{\mathcal{G}}\mathbf{B}\backslash\mathbf{C}_1|\mathbf{C}_1$ and $\mathbf{A}\perp\!\!\!\perp_{\mathcal{G}}\mathbf{B}\backslash\mathbf{C}_2|\mathbf{C}_2$ if and only if $\mathbf{A}\perp\!\!\!\perp_{\mathcal{G}}(\mathbf{R},\mathbf{W_2})\,|\mathbf{W_1},\mathbf{I}$ and $\mathbf{A}\perp\!\!\!\perp_{\mathcal{G}}(\mathbf{R},\mathbf{W_1})\,|\mathbf{W_2},\mathbf{I}$.The weak union axiom implies that

$$\mathbf{A}\perp\!\!\!\perp_{\mathcal{G}}\mathbf{R}|\,(\mathbf{W_1},\mathbf{W_2}),\mathbf{I} \tag{29}$$

The decomposition axiom implies that

$$\mathbf{A}\perp\!\!\!\perp_{\mathcal{G}}\mathbf{W_2}|\mathbf{W_1},\mathbf{I}\text{ and }\mathbf{A}\perp\!\!\!\perp_{\mathcal{G}}\mathbf{W_1}|\mathbf{W_2},\mathbf{I}. \tag{30}$$

Next, it follows from (30) and the intersection axiom that

$$\mathbf{A}\perp\!\!\!\perp_{\mathcal{G}}(\mathbf{W_1},\mathbf{W_2})\,|\mathbf{I} \tag{31}$$

Finally, from (29) and (31), the contraction axiom implies that $\mathbf{A}\perp\!\!\!\perp_{\mathcal{G}}(\mathbf{R},\mathbf{W_1},\mathbf{W_2})\,|\mathbf{I}$ or, equivalently, that $\mathbf{A}\perp\!\!\!\perp_{\mathcal{G}}\mathbf{B}\backslash\mathbf{I}|\mathbf{I}$. If $\mathbf{C}_1$ and $\mathbf{C}_2$ are distinct then $\mathbf{I}$ is a strict subset of $\mathbf{C}_1$ and $\mathbf{C}_2$ which cannot happen because $\mathbf{C}_1$ and $\mathbf{C}_2$ were minimal sets $\mathbf{C}'$ with the property that $\mathbf{A}\perp\!\!\!\perp_{\mathcal{G}}\mathbf{B}\backslash\mathbf{C}'|\mathbf{C}'$. ■

Proof [Proof of Theorem 9]

Proof of part (1). From $A\perp\!\!\!\perp_{\mathcal{G}}\mathbf{O}\setminus\mathbf{O}_{min}\mid\mathbf{O}_{min}$, we have that $\pi_a(\mathbf{O},P)=\pi_a(\mathbf{O}_{min},P)$. Then, for all $P\in\mathcal{M}(\mathcal{G}), E_P\left[\frac{I_a(A)Y}{\pi_a(\mathbf{O}_{min},P)}\right]=E_P\left[\frac{I_a(A)Y}{\pi_a(\mathbf{O},P)}\right]=\chi_a(P;\mathcal{G})$ where the last equality follows because, since A is a point intervention, $\mathbf{O}$ is an adjustment set. This shows that $\mathbf{O}_{min}$ is an adjustment set

Proof of part (2). In our proof of part (2) we will invoke at several places the following property.

Property (O): For any $O \in \mathbf{O}$ there exists a directed path from O to Y such that for any adjustment set $\mathbf{Z}$ relative to (A, Y) in $\mathcal{G}$, the path does not intersect the nodes in $\mathbf{Z}$ other than, at most, at the node O.

The proof of Property (O) is immediate because by definition of $\mathbf{O}$, O is the parent of a node in $\mathrm{cn}(A, Y; \mathcal{G})$. If such node is Y then the assertion holds trivially for the path $O \to Y$. Otherwise, for any node M in $\mathrm{cn}(A, Y; \mathcal{G}) \setminus \{Y\}$ there exists a directed path from M to Y that intersects solely nodes in $\mathrm{cn}(A, Y; \mathcal{G})$. The assertion then holds for such path because for any adjustment set $\mathbf{Z}$ it holds that $\mathbf{Z} \cap \mathrm{cn}(A, Y; \mathcal{G}) = \emptyset$.

Turn now to the proof of $A \perp\!\!\!\perp_{\mathcal{G}} [\mathbf{O}_{\min} \backslash \mathbf{Z}_{\min}] \mid \mathbf{Z}_{\min}$.Suppose there existed $O \in \mathbf{O}_{\min} \backslash \mathbf{Z}_{\min}$ such that O is not d-separated from A given $\mathbf{Z}_{\min}$ in $\mathcal{G}$. Let α denote the path between A and O that is open given $\mathbf{Z}_{\min}$. By Property (O) there exists a directed path, say φ, between O and Y that is open given $\mathbf{Z}_{\min}$. Then, the path obtained by concatenating α with φ is a non-causal path between A and Y that is open given $\mathbf{Z}_{\min}$, which is impossible because $\mathbf{Z}_{\min}$ is an adjustment set.

Turn now to the proof of

$$Y \perp\!\!\!\perp_{\mathcal{G}} [\mathbf{Z}_{min} \setminus \mathbf{O}_{min}] \mid \mathbf{O}_{min}, A. \tag{32}$$

We will show it by contradiction. Assume there exists $X^* \in \mathbf{Z}_{min} \backslash \mathbf{O}_{min}$ such that $Y \not\perp\!\!\!\perp_{\mathcal{G}} X^* \mid \mathbf{O}_{min}, A$. By $X^* \in \mathbf{Z}_{min}$, Theorem 5 from Shpitser et al. (2010), implies that

$$X^* \notin \mathrm{de}_{\mathcal{G}}(A). \tag{33}$$

Also, by Shpitser et al. (2010), we have

$$X^* \notin \mathrm{forb}\,(A, Y; \mathcal{G})\,. \tag{34}$$

Let η^* be the path between X^* and Y that is open when we condition on $(\mathbf{O}_{min}, A)$. We will first show that η^* must intersect a vertex in $\mathbf{O} \backslash \mathbf{O}_{min}$. So, if $\mathbf{O} \backslash \mathbf{O}_{min} = \emptyset$, this result already shows (32) . To show that η^* must intersect a vertex in $\mathbf{O} \backslash \mathbf{O}_{min}$ we first note that η^* must be of the form $X^* - \circ ... \circ \to Y$. The justification for why the last edge in η^* must point into Y is as follows. Suppose the edge pointed out of Y. Then, since by (34) X^* cannot be a descendant of Y, the path η^* would have to intersect a vertex that would be both a descendant of Y and a collider in η^*, and either such vertex or any of its descendants would have to be in the conditioning set $\mathbf{O}_{min} \cup \{A\}$ so as to yield the path η^* open. But this is impossible because neither A can be a descendant of Y nor can any element of $\mathbf{O}_{min}$, by the very definition of $\mathbf{O}_{min}$.

Next we note that, by definition of $\mathbf{O}$, in the edge $\circ \to Y$ the vertex $\circ$ is in the set $\mathbf{O} \cup \mathbf{M}$ where $\mathbf{M} \equiv \mathrm{cn}\,(A, Y; \mathcal{G}) \setminus \{Y\}$. If the vertex is in $\mathbf{O}$ then it must be in $\mathbf{O} \backslash \mathbf{O}_{min}$ because the path η^* is open when conditioning on $(\mathbf{O}_{min}, A)$, therefore proving the assertion that η^* intersects $\mathbf{O} \backslash \mathbf{O}_{min}$. If the vertex is in $\mathbf{M}$, then the next edge in the path must be of the form $X^* - \circ ... \circ \to M_k \to Y$ for some $M_k \in \mathbf{M}$. The justification for why the edge $\circ \to M_k$ points into M_k is along the same lines as before. Specifically, if the edge pointed out of M_k then, by virtue of X^* not being a descendant of M_k, then the path η^* would have to intersect a vertex that would be both a descendant of M_k and a collider in η^*, and either

such vertex or any of its descendants would have to be in the conditioning set $\mathbf{O}_{min} \cup \{A\}$. But this is impossible because neither A can be a descendant of M_k nor can any element of $\mathbf{O}_{min}$, by the very definition of $\mathbf{O}_{min}$.

By the same argument as above, in the edge $\circ \to M_k$ the vertex $\circ$ is in the set $\mathbf{O} \cup \mathbf{M}$. If the vertex is in $\mathbf{O}$ then it must be in $\mathbf{O} \backslash \mathbf{O}_{min}$ because the path η^* is open when conditioning on $(\mathbf{O}_{min}, A)$, therefore proving the assertion that η^* intersects $\mathbf{O} \backslash \mathbf{O}_{min}$. If the vertex is a, say M_j, in $\mathbf{M}$ then reasoning as above, the path η^* must be of the form $X^* - \circ ... \circ \to M_j \to M_k \to Y$. Continuing in the same fashion, we arrive at the conclusion that either any of the vertices $\circ$ are in $\mathbf{O} \backslash \mathbf{O}_{min}$ or otherwise, the path is of the form $X^* \to M_r \to M_l \to ... \to M_j \to M_k \to Y$. In the latter case, $X^* \in \mathbf{O} \backslash \mathbf{O}_{min}$ which therefore concludes the proof that the path η^* intersects $\mathbf{O} \backslash \mathbf{O}_{min}$.

Let $O^* \in \mathbf{O} \setminus \mathbf{O}_{min}$ be the element of $\mathbf{O} \setminus \mathbf{O}_{min}$ that is closest to Y in the path η^*, that is, such that the subpath of η^* between O^* and Y does not intersect any other vertex of $\mathbf{O} \setminus \mathbf{O}_{min}$. Let $D_1^*, \ldots, D_k^*$ be the colliders on η^*, with D_1^* the one closest to X^* in η^*, D_2^* the one second closest to X^* and so on. For each j there exists a descendant of D_j^* that is an element of $(\mathbf{O}_{min}, A)$. Furthermore, if there exists a directed path between D_j^* and A, this path necessarily has to intersect an element of $\mathbf{O}_{min}$ for suppose this was not the case. Then, take j^* to be the largest j such that there exists a directed path between D_j^* and A that does not intersect $\mathbf{O}_{min}$. Then the path $A \leftarrow ... \leftarrow D_{j^*}^* \leftarrow ... - O^*$ is open given $\mathbf{O}_{min}$, which contradicts $O^* \in \mathbf{O} \backslash \mathbf{O}_{min}$. We therefore conclude that η^* is open by conditioning just on $\mathbf{O}_{min}$.

From the nodes in $\mathbf{Z}_{min} \backslash \mathbf{O}_{min}$ that intersect η^*, let W^* be the closest one to O^* in the path η^*, possibly $W^* = O^*$. Consider now the subpath α^* of η^* between W^* and Y. Because η^* is open by conditioning on $\mathbf{O}_{min}$, so is α^*. The path α^* has one of the following two forms

$$W^* \to \circ - ... \to \underbrace{\circ}_{\equiv C_1^*} \leftarrow ... \to \underbrace{\circ}_{\equiv C_2^*} \leftarrow .... \to \underbrace{\circ}_{\equiv C_r^*} \leftarrow ... - O^* \to M_{u_1} \to M_{u_2} ... \to M_{u_t} \to Y \tag{35}$$

or

$$W^* \leftarrow \circ - ... \to \underbrace{\circ}_{\equiv C_1^*} \leftarrow ... \to \underbrace{\circ}_{\equiv C_2^*} \leftarrow .... \to \underbrace{\circ}_{\equiv C_r^*} \leftarrow ... - O^* \to M_{u_1} \to M_{u_2} ... \to M_{u_t} \to Y. \tag{36}$$

where the set of colliders $\{C_1^*, ..., C_r^*\}$ is included in $\{D_1^*, ..., D_k^*\}$ and can possibly be empty, and the set $\{M_{u_1}, M_{u_2}, ..., M_{u_t}\}$ is included in $\mathbf{M}$ and can also possibly be empty.

Next, let

$$\Delta \equiv \{\delta : \delta \text{ is a path between } W^* \text{ and } A \text{ that is open given } \mathbf{Z}_{min} \backslash \{W^*\}\}.$$

Lemma 28 in Section A.4 implies that Δ is not empty. Any path δ in Δ has one of the following forms:

a) δ is a directed path from W^* to A : $W^* \to \circ \to \circ ... \circ \to A$

b) δ has one and only one fork: $W^* \leftarrow \circ ... \circ \leftarrow \circ \to \circ ... \circ \to A$.

c) δ has at least one collider and the first edge points out of W^*:

$$W^* \to \circ - ... \to \underbrace{\circ}_{\equiv H_1^*} \leftarrow ... \to \underbrace{\circ}_{\equiv H_2^*} \leftarrow .... \to \underbrace{\circ}_{\equiv H_s^*} \leftarrow ... - A.$$

d) δ has at least one collider and the first edge points into W^*:

$$W^* \leftarrow \circ - ... \rightarrow \underbrace{\circ}_{\equiv H_1^*} \leftarrow ... \rightarrow \underbrace{\circ}_{\equiv H_2^*} \leftarrow .... \rightarrow \underbrace{\circ}_{\equiv H_s^*} \leftarrow ... - A.$$

Moreover, we can assume without loss of generality that W^* appears only once in the path δ. Note that $\delta \in \Delta$ cannot be a directed path from A to W^* because $W^* \in \mathbf{Z}_{min}$ and by Theorem 5 from Shpitser et al. (2010) we have that $W^* \notin \text{de}_{\mathcal{G}}(A)$.

We will show that neither of the forms (35) or (36) for the path α^* are possible by showing that if α^* was of one such form then it would imply that Δ is empty. Henceforth, assume that α^* takes one of the forms (35) or (36) . Below we will show the following claims.

Claim (i). $\forall\ \delta \in \Delta$ with form (a) or (b), δ is open given $\mathbf{O}_{min}$.

Claim (ii). If $\exists \delta \in \Delta$ with form (b) or (d) then the path α^* cannot be of the form (36).

Claim (iii). Every $\delta \in \Delta$ of the form (c) or (d) is blocked given $\mathbf{O}_{min}$.

Proof of Claim (i). Let δ have form (a) or (b). Then no node in $\mathbf{O}_{min} \cap \mathbf{Z}_{min}$ intersects δ, for if it did, the path would be blocked by $\mathbf{Z}_{min} \backslash \{W^*\}$. On the other hand, suppose the path δ intersected a node O^{**} in $\mathbf{O}_{min} \backslash \mathbf{Z}_{min}$. Let ξ be the subpath of δ between O^{**} and A. The path ξ is open given $\mathbf{Z}_{min}$. By Property (O) there exists a directed path, say φ, from O^{**} to Y that does not intersect $\mathbf{Z}_{min}$. Then, the path obtained by concatenating ξ with φ is a non-causal path between A and Y that is open given $\mathbf{Z}_{min}$. This contradicts the assumption that $\mathbf{Z}_{min}$ is an adjustment set. This concludes the proof of Claim (i).

Proof of Claim (ii). Suppose that there exists a path $\delta \in \Delta$ with form (b) or (d). We will prove by contradiction that there cannot be any path α^* of the form (36) that is open when by conditioning on $\mathbf{O}_{min}$. Suppose there existed one such path α^*. Suppose first that there are no colliders in α^*, that is, there exist no nodes C_j^*. The path α^* does not intersect any element of $\mathbf{Z}_{min} \backslash \mathbf{O}_{min}$ other than at the node W^* because, by definition, W^* was chosen to be the closest element in $\mathbf{Z}_{min} \backslash \mathbf{O}_{min}$ to O^*. On the other hand, since the path α^* is open by conditioning on $\mathbf{O}_{min}$, then α^* cannot intersect any element of $\mathbf{O}_{min}$. Then, α^* is open given $\mathbf{Z}_{min} \backslash W^*$. Take now the path $\delta \in \Delta$ with form (b) or (d) and concatenate it with the path α^*. The concatenated path is a non-causal path between A and Y which is open given $\mathbf{Z}_{min}$ because W^* is a collider in the path and $W^* \in \mathbf{Z}_{min}$. This is impossible because $\mathbf{Z}_{min}$ is an adjustment set. We therefore conclude if a path α^* exists, then the set of colliders $\{C_1^*, ..., C_k^*\}$ is not empty. Furthermore, at least one of the colliders is not an ancestor of any node in $\mathbf{Z}_{min}$, for if all $C_1^*, ..., C_k^*$ were ancestors of some node in $\mathbf{Z}_{min}$, then again the path α^* would be open given $\mathbf{Z}_{min} \backslash W^*$ and consequently, the concatenated path between a path δ of the form (b) or (d) with the path α^* would be a non-causal path between A and Y that is open given $\mathbf{Z}_{min}$, contradicting the assumption that $\mathbf{Z}_{min}$ is an adjustment set. Take the smallest j, say j', such that the collider C_j^* is not an ancestor of $\mathbf{Z}_{min}$. Because the path α^* is open by conditioning on $\mathbf{O}_{min}$, then there exists $O^{**} \in \mathbf{O}_{min} \backslash \mathbf{Z}_{min}$ such that $C_{j'}^*$ is an ancestor of O^{**} so that either there exists a directed path, say λ, from $C_{j'}^*$ to O^{**} or $C_{j'}^* = O^{**}$. Now, by Property (O) there exists a directed path, say φ, from O^{**} to Y that is open given $\mathbf{Z}_{min}$. Now, consider the path that concatenates a path $\delta \in \Delta$ with form (b) or (d), with the subpath of α^* between W^* and $C_{j'}^*$, next concatenates with λ if $C_{j'}^* \neq O^{**}$ and finally concatenates with φ. Such path is

a non-causal path between A and Y that is open given $\mathbf{Z}_{min}$ which is impossible because $\mathbf{Z}_{min}$ is an adjustment set. This concludes the proof of the Claim (ii)

Proof of Claim (iii). Suppose that $\delta \in \Delta$ is of the form (c) and that δ is open given $\mathbf{O}_{min}$. Then concatenating δ with α^* we obtain a non-causal path between A and Y that is open given $\mathbf{O}_{min}$ because W^* is not a collider on this path. This contradicts the fact that $\mathbf{O}_{min}$ is an adjustment set.

Suppose now that $\delta \in \Delta$ is of the form (d) and is open given $\mathbf{O}_{min}$. By Claim (ii), the path α^* has to be of the form (35). Then concatenating δ with α^* we once again obtain a path between A and Y that is open given $\mathbf{O}_{min}$ arriving at a contradiction. This concludes the proof of Claim (iii).

We will now argue that Δ must be empty by showing that Claims (i), (ii) and (iii) imply that if $\delta \in \Delta$, then δ cannot take any of the forms (a), (b), (c) or (d).

(I) Proof that $\delta \in \Delta$ cannot take the form (a). Suppose there exists $\delta \in \Delta$ with the form (a). Then, invoking Claim (i), we conclude that the path γ between O^* and A formed by concatenating the path δ between W^* and A and the subpath of α^* between O^* and W^* is a path between O^* and A that is open given $\mathbf{O}_{min}$. This is impossible because the existence of such path γ contradicts the assertion that $O^* \in \mathbf{O} \setminus \mathbf{O}_{min}$.

(II) Proof that $\delta \in \Delta$ cannot take the form (b). Suppose there exists $\delta \in \Delta$ with the form (b). Then invoking Claim (ii), the path α^* has to be of the form (35). By Claim (i), δ is open given $\mathbf{O}_{min}$. On the other hand, α^* is open given $\mathbf{O}_{min}$. Concatenating δ with α^* we form a path, say π, that is open given $\mathbf{O}_{min}$, since W^* is not a collider on π. This is impossible because the existence of such path π contradicts the fact that $O^* \in \mathbf{O} \setminus \mathbf{O}_{min}$.

(III) Proof that $\delta \in \Delta$ can take neither the form (c) nor the form (d). Suppose that there exists a $\delta \in \Delta$ of the form (c) or (d). By Claim (iii), δ is blocked by conditioning on $\mathbf{O}_{min}$. Furthermore, by definition of Δ, the path is open when conditioning on $\mathbf{Z}_{min} \backslash W^*$. Then, one of the following happens:

(III.a) the path δ intersects a node $O^{**} \in \mathbf{O}_{min} \backslash \mathbf{Z}_{min}$ that is not a collider in the path, or

(III.b) the property (III.a) does not hold and there exists a non-empty subset, say $\mathcal{H} \equiv \left\{H^*_{j_1}, \ldots, H^*_{j_l}\right\}$, of the collider set $\{H^*_1, ..., H^*_s\}$ such that each $H^*_{j_u}$ is an ancestor in $\mathcal{G}$ of a node in $\mathbf{Z}_{min} \backslash W^*$ but is not an ancestor of a node in $\mathbf{O}_{min}$.

We will show by contradiction that both (III.a) and (III.b) are impossible.

Suppose first that (III.a) holds. Let ϕ be the subpath in δ between O^{**} and A. The path ϕ is open given $\mathbf{Z}_{min}$ because the path δ is open given $\mathbf{Z}_{min} \backslash W^*$. Let ν a directed path between O^{**} and Y that does not intersect $\mathbf{Z}_{min}$, which exists by Property (O). The path between A and Y obtained by concatenating ν with ϕ is a non-causal path between A and Y that is open given $\mathbf{Z}_{min}$. This is impossible because $\mathbf{Z}_{min}$ is an adjustment set.

Suppose next that (III.b) holds. Let $\mathcal{H} = \left\{H^*_{j_1}, ..., H^*_{j_l}\right\}$ be the maximal subset of the collider set $\{H^*_1, ..., H^*_s\}$ such that each $H^*_{j_u}$ is an ancestor in $\mathcal{G}$ of a node in $\mathbf{Z}_{min} \backslash W^*$ but is not an ancestor of a node in $\mathbf{O}_{min}$. Assume without loss of generality that $j_1 < j_2 < \cdots < j_l$ so that $H^*_{j_l}$ is the closest node in $\mathcal{H}$ to A in the path δ. Then, the subpath of δ, say ζ_{j_l}, between $H^*_{j_l}$ and A is open given $\mathbf{O}_{min}$.

We will show next that if (III.b) holds then

$$\text{for each } u \text{ in } \{1, ..., l\} \text{ there exists } O^*_{j_u} \in \mathbf{O} \backslash \mathbf{O}_{min} \text{ such that } H^*_{j_u} \not\perp\!\!\!\perp_{\mathcal{G}} O^*_{j_u} | \mathbf{O}_{min}. \quad (37)$$

However, (37) leads to a contradiction. To see this, let $\nu^*_{j_l}$ be a directed path between $O^*_{j_l}$ and Y that does not intersect $\mathbf{O}_{min}$, which exists by Property (O). Let $\gamma^*_{j_l}$ denote the path between $H^*_{j_l}$ and $O^*_{j_l}$ which is open by conditioning on $\mathbf{O}_{min}$ (which exists by (37)). Then, the path obtained by concatenating the paths $\nu^*_{j_l}$ with $\gamma^*_{j_l}$ and with ζ_{j_l} is a non-causal path between A and Y that is open by conditioning on $\mathbf{O}_{min}$. This is impossible because $\mathbf{O}_{min}$ is an adjustment set. The proof of part (2) of the theorem is then finished if we show that (III.b) implies (37). We will show this by induction in u. Suppose first that $u = 1$. By definition of the set $\mathcal{H}$, either $H^*_{j_1} \equiv Z^*_{u=1,1} \in \mathbf{Z}_{min} \backslash \{\mathbf{O}_{min}, W^*\}$ or there exists a node $Z^*_{u=1,1} \in \mathbf{Z}_{min} \backslash \{\mathbf{O}_{min}, W^*\}$ such that there exists a directed path from $H^*_{j_1}$ to $Z^*_{u=1,1}$ that does not intersect any other element of $\mathbf{Z}_{min} \backslash \{\mathbf{O}_{min}, W^*\}$. Now, because $Z^*_{u=1,1} \in \mathbf{Z}_{min}$ and $\mathbf{Z}_{min}$ is a minimal adjustment set, then there exists a non-causal path θ_1 between A and Y such that θ_1 is open by conditioning on $\mathbf{Z}_{min} \backslash Z^*_{u=1,1}$, and the path θ_1 is closed by conditioning on $\mathbf{Z}_{min}$.The path θ_1 must then intersect $Z^*_{u=1,1}$ and $Z^*_{u=1,1}$ must be a non-collider vertex in the path. Now, define $\tau_1 \equiv$ subpath of θ_1 between A and $Z^*_{u=1,1}$, and $\kappa_1 \equiv$ subpath of θ_1 between $Z^*_{u=1,1}$ and Y. Because $Z^*_{u=1,1}$ is a non-collider in the path θ_1, then in at least one of the subpaths τ_1 or κ_1, the edge with vertex $Z^*_{u=1,1}$ is pointing out of $Z^*_{u=1,1}$. Furthermore, because θ_1 is open by conditioning on $\mathbf{Z}_{min} \backslash Z^*_{u=1,1}$, so are τ_1 and κ_1. We will show next that either

$$\text{(37) holds for } u = 1 \text{ or } \exists \text{ a vertex } Z^*_{u=1,2} \text{ in } \mathbf{Z}_{min} \backslash \left[\mathbf{O}_{min} \cup \left\{Z^*_{u=1,1}\right\}\right] \text{ such that } Z^*_{u=1,2} \text{ is a descendant of } Z^*_{u=1,1} \tag{38}$$

Suppose first that the edge with vertex $Z^*_{u=1,1}$ in τ_1 points out of $Z^*_{u=1,1}$. We will now show that τ_1 cannot be a directed path from $Z^*_{u=1,1}$ to A. Suppose τ_1 was a directed path. Then τ_1 cannot intersect any vertex of $\mathbf{O}_{min}$, because $H^*_{j_1}$), and hence $Z^*_{u=1,1}$, is not ancestor of any vertex in $\mathbf{O}_{min}$. We therefore conclude that if τ_1 is a directed path between $Z^*_{u=1,1}$ and A, then it must be open by conditioning on $\mathbf{O}_{min}$. Now, let λ be the subpath of δ^* between W^* and $H^*_{j_1}$. By definition of $H^*_{j_1}, \lambda$ is open given $\mathbf{O}_{min}$. Let ρ be the directed path between $H^*_{j_1}$ and $Z^*_{u=1,1}$ if $H^*_{j_1} \neq Z^*_{u=1,1}$, otherwise let ρ denote the degenerate path consisting of just the vertex $H^*_{j_1}$. Let

$$\beta \equiv \text{ the path between } A \text{ and } Y \text{ obtained by concatenating } \tau_1 \text{ with } \rho \text{ with } \lambda \text{ with } \alpha^*.$$

Because all the paths τ_1, ρ, λ and α^* are open given $\mathbf{O}_{min}$ and because none of the vertices $W^*, H^*_{j_1}$ and $Z^*_{u=1,1}$ are in $\mathbf{O}_{min}$, and none are colliders in the path β, then the path β is open given $\mathbf{O}_{min}$. This is impossible because $\mathbf{O}_{min}$ is an adjustment set. We therefore conclude that τ_1 cannot be a directed path between $Z^*_{u=1,1}$ and A. Therefore, τ_1 must intersect a collider. Any collider in the path τ_1 must be an ancestor of a node in the set $\mathbf{Z}_{min} \backslash \left\{Z^*_{u=1,1}\right\}$ because τ_1 is open given $\mathbf{Z}_{min} \backslash \left\{Z^*_{u=1,1}\right\}$. Furthermore, the collider in τ_1 that is closest to $Z^*_{u=1,1}$ cannot be an ancestor of any element of $\mathbf{O}_{min}$, because if it was, then $Z^*_{u=1,1}$ and consequently $H^*_{j_1}$ would be an ancestor of a vertex in $\mathbf{O}_{min}$, which is not possible by the definition of the set $\mathcal{H}$. We therefore conclude that there exists a vertex, say $Z^*_{u=1,2}$, in $\mathbf{Z}_{min} \backslash \left[\mathbf{O}_{min} \cup \left\{Z^*_{u=1,1}\right\}\right]$ such that $Z^*_{u=1,2}$ is a descendant of $Z^*_{u=1,1}$, thus showing (38).

Next suppose that the edge with vertex $Z^*_{u=1,1}$ in κ_1 points out of $Z^*_{u=1,1}$. If there exists a directed path between $Z^*_{u=1,1}$ and Y, then this path necessarily has to intersect an element $O^*_{j_1} \in \mathbf{O}$. The vertex $O^*_{j_1}$ cannot be in $\mathbf{O}_{min}$ because if it were, then $H^*_{j_1}$ would be an

ancestor of an element of $\mathbf{O}_{min}$, which is impossible by the definition of the set $\mathcal{H}$. Then, if there exists a directed path between $Z^*_{u=1,1}$ and Y, the assertion (38) holds. Now, suppose that there exists no directed path between $Z^*_{u=1,1}$ and Y. Then, the path κ_1 must intersect a collider. Because κ_1 is open given $\mathbf{Z}_{min} \setminus \left\{Z^*_{u=1,1}\right\}$ and because $Z^*_{u=1,1}$ cannot be the ancestor of any vertex in $\mathbf{O}_{min}$, then we reason exactly as before, and conclude that there exists a $Z^*_{u=1,2}$, in $\mathbf{Z}_{min} \setminus \left[\mathbf{O}_{min} \cup \left\{Z^*_{u=1,1}\right\}\right]$ such that $Z^*_{u=1,2}$ is a descendant of $Z^*_{u=1,1}$, thus proving (38) .

Next, suppose (38) holds because there exists a vertex $Z^*_{u=1,2}$ in $\mathbf{Z}_{min} \setminus \left[\mathbf{O}_{min} \cup \left\{Z^*_{u=1,1}\right\}\right]$ such that $Z^*_{u=1,2}$ is a descendant of $Z^*_{u=1,1}$. We can now reason exactly as we did for $Z^*_{u=1,1}$ and conclude that

$$\text{(37) holds for } u=1 \text{ or } \exists \text{ a vertex } Z^*_{u=1,3} \text{ in } \mathbf{Z}_{min} \setminus \left[\mathbf{O}_{min} \cup \left\{Z^*_{u=1,1}, Z^*_{u=1,2}\right\}\right] \text{ such that } Z^*_{u=1,3} \text{ is a descendant of } Z^*_{u=1,2} \tag{39}$$

Continuing in this fashion until depleting the set of vertices in $\mathbf{Z}_{min}$ we arrive at the conclussion that (37) holds for $u = 1$.

Suppose now that (37) holds for $u = 1, ..., t-1$ with $t \leq l$. We will show that it holds for $u = t$. Let $Z^*_{u=t,1} \in \mathbf{Z}_{min} \setminus \mathbf{O}_{min}$ be a descendant of $H^*_{j_t}$ which exists by the definition of $\mathcal{H}$. Let θ_t be a path that is open given $\mathbf{Z}_{min} \setminus Z^*_{u=t,1}$ but closed given $\mathbf{Z}_{min}$. Reasoning as before, the path θ_t must intersect $Z^*_{u=t,1}$ and $Z^*_{u=t,1}$ cannot be a collider in the path. Then, partitioning θ_t as (τ_t, κ_t) where $\tau_t \equiv$ subpath of θ_t between A and $Z^*_{u=t,1}$ and $\kappa_t \equiv$ subpath of θ_t between $Z^*_{u=t,1}$ and Y we know that in at least one of τ_t or κ_t the edge with one endpoint equal to $Z^*_{u=t,1}$ must point out of $Z^*_{u=t,1}$. Furthermore, both τ_t and κ_t are open given $\mathbf{Z}_{min} \setminus Z^*_{u=t,1}$. We will show that

$$\text{(37) holds for } u=t \text{ or } \exists \text{ a vertex } Z^*_{u=t,2} \text{ in } \mathbf{Z}_{min} \setminus \left[\mathbf{O}_{min} \cup \left\{Z^*_{u=t,1}\right\}\right] \text{ such that } Z^*_{u=t,2} \text{ is a descendant of } Z^*_{u=t,1} \tag{40}$$

Suppose the edge with one endpoint equal to $Z^*_{u=t,1}$ in τ_t points out of $Z^*_{u=t,1}$. We will show that τ_t cannot be a directed path from $Z^*_{u=t,1}$ to A. As we reasoned for τ_1 above, if τ_t was directed it could not intersect any element of $\mathbf{O}_{min}$, for if it did, then such element of $\mathbf{O}_{min}$ would be a descendant of $H^*_{j_t}$ which is impossible by the definition of the set $\mathcal{H}$. So, if a directed path between $Z^*_{u=t,1}$ and A exists, then it must be open given $\mathbf{O}_{min}$. Now, by the inductive hypothesis, we know that there exists $O^*_{j_{t-1}} \in \mathbf{O} \setminus \mathbf{O}_{min}$ such that $O^*_{j_{t-1}}$ is a descendant of $H^*_{j_{t-1}}$. Because, by definition of $\mathcal{H}$, $H^*_{j_{t-1}}$ cannot be an ancestor of any vertex in $\mathbf{O}_{min}$, then we conclude that there exists a directed path, say σ, from $H^*_{j_{t-1}}$ and $O^*_{j_{t-1}}$ that is open by conditioning on $\mathbf{O}_{min}$. Let $\lambda_{j_{t-1}}$ be the subpath of δ between $H^*_{j_{t-1}}$ and $H^*_{j_t}$. The path $\lambda_{j_{t-1}}$ is open by conditioning on $\mathbf{O}_{min}$ because we have assumed that δ does not intersect any node of $\mathbf{O}_{min}$ that is a non-collider in the path, and by the definition of $H^*_{j_{t-1}}$ and $H^*_{j_t}$, if in the path $\lambda_{j_{t-1}}$ there are colliders, each of these colliders must be ancestors of $\mathbf{O}_{min}$. Let $\rho_{j_{t-1}}$ be the directed path between $H^*_{j_t}$ and $Z^*_{u=t,1}$ if $H^*_{j_t} \neq Z^*_{u=t,1}$, otherwise let $\rho_{j_{t-1}}$ denote the degenerate path consisting of just the vertex $H^*_{j_t}$. Note that $\rho_{j_{t-1}}$ is open given $\mathbf{O}_{min}$ because $H^*_{j_t}$ is not an ancestor of any vertex in $\mathbf{O}_{min}$. Let $\beta_{j_t} =$ the path between A and $O^*_{j_{t-1}}$ obtained by concatenating τ_t with $\rho_{j_{t-1}}$ with $\lambda_{j_{t-1}}$ with σ. Because all the paths τ_t, $\rho_{j_{t-1}}, \lambda_{j_{t-1}}$ and σ are open given $\mathbf{O}_{min}$ and because none of the vertices $H^*_{j_{t-1}}, H^*_{j_t}$

and $Z^*_{u=t,1}$ are in $\mathbf{O}_{min}$, and none are colliders in the path β_{j_t}, then the path β_{j_t} is open given $\mathbf{O}_{min}$. This is impossible because by definition of $\mathbf{O}_{min}$, $O^*_{j_{t-1}}$ is d-separated from A given $\mathbf{O}_{min}$. We therefore conclude that τ_t cannot be a directed path between $Z^*_{u=t,1}$ and A. Therefore, τ_t must intersect a collider. Any collider in the path τ_t must be an ancestor of a node in the set $\mathbf{Z}_{min} \backslash \left\{Z^*_{u=t,1}\right\}$ because τ_t is open given $\mathbf{Z}_{min} \backslash \left\{Z^*_{u=t,1}\right\}$. Furthermore, the collider in τ_t that is closest to $Z^*_{u=t,1}$ cannot be an ancestor of any element of $\mathbf{O}_{min}$, because if it was, then $Z^*_{u=t,1}$ and consequently $H^*_{j_t}$ would be an ancestor of a vertex in $\mathbf{O}_{min}$, which is not possible by the definition of the set $\mathcal{H}$. We therefore conclude that there exists a vertex, say $Z^*_{u=t,2}$, in $\mathbf{Z}_{min} \backslash \left[\mathbf{O}_{min} \cup \left\{Z^*_{u=t,1}\right\}\right]$ such that $Z^*_{u=t,2}$ is a descendant of $Z^*_{u=t,1}$, thus showing (40) holds if the edge with one endpoint equal to $Z^*_{u=t,1}$ in τ_t points out of $Z^*_{u=t,1}$.

Suppose next that the edge with one endpoint equal to $Z^*_{u=t,1}$ in κ_t points out of $Z^*_{u=t,1}$. If there exists a directed path between $Z^*_{u=t,1}$ and Y, then this path necessarily has to intersect an element $O^*_{j_t} \in \mathbf{O}$. The vertex $O^*_{j_t}$ cannot be in $\mathbf{O}_{min}$ because if it were, then $H^*_{j_t}$ would be an ancestor of an element of $\mathbf{O}_{min}$, which is impossible by the definition of the set $\mathcal{H}$. Then, if there exists a directed path between $Z^*_{u=t,1}$ and Y, the assertion (40) holds. Now, suppose that there exists no directed path between $Z^*_{u=t,1}$ and Y. Then, the path κ_t must intersect a collider. Because κ_t is open given $\mathbf{Z}_{min} \backslash \left\{Z^*_{u=t,1}\right\}$ and because $Z^*_{u=t,1}$ cannot be the ancestor of any vertex in $\mathbf{O}_{min}$, then we reason exactly as before, and conclude that there exists a $Z^*_{u=t,2}$, in $\mathbf{Z}_{min} \backslash \left[\mathbf{O}_{min} \cup \left\{Z^*_{u=t,1}\right\}\right]$ such that $Z^*_{u=t,2}$ is a descendant of $Z^*_{u=t,1}$.

Next, because $Z^*_{u=t,2}$ is in $\mathbf{Z}_{min} \backslash \left[\mathbf{O}_{min} \cup \left\{Z^*_{u=t,1}\right\}\right]$ and is a descendant of $Z^*_{u=t,1}$, we can reason exactly as we did for $Z^*_{u=t,1}$ and conclude that

(37) holds for $u = t$ or $\exists$ a vertex $Z^*_{u=t,3}$ in $\mathbf{Z}_{min} \backslash \left[\mathbf{O}_{min} \cup \left\{Z^*_{u=t,1}, Z^*_{u=t,2}\right\}\right]$ such that $Z^*_{u=t,3}$ is a descendant of $Z^*_{u=t,2}$.

Continuing in this fashion until depleting the set of vertices in $\mathbf{Z}_{min}$ we arrive at the conclussion that (37) holds for $u = t$. This concludes the proof of the part (2).

Proof of part (3). Suppose there existed a minimal adjustment set $\mathbf{Z}_{min}$ that contained a vertex $O \in \mathbf{O} \backslash \mathbf{O}_{min}$. Then $O \in \mathbf{O}$ and $O \in \mathbf{Z}_{min} \backslash \mathbf{O}_{min}$. Part (2) of this Theorem then implies $Y \perp\!\!\!\perp_{\mathcal{G}} O \mid \mathbf{O}_{min}, A$. This is impossible because by Property (O) there exists a directed path from O to Y that does not intersect $\mathbf{O}_{min}$. The path also does not intersect A. Consequently, by virtue of being a directed path, the path is open given $\mathbf{O}_{min}$ and A. This concludes the proof of the Theorem. ■

Proof [Proof of Lemma 11] First note that for $k \in \{0, \ldots, p\}$,

$$\begin{aligned} \pi_{a_k}\left(\overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k; P\right) \equiv P\left(A_k = a_k | \overline{\mathbf{A}}_{k-1} = \overline{\mathbf{a}}_{k-1}, \overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k\right) &= P\left(A_k = a_k | \overline{\mathbf{A}}_{k-1} = \overline{\mathbf{a}}_{k-1}, \overline{\mathbf{B}}_k\right) \\ &\equiv \pi_{a_k}\left(\overline{\mathbf{B}}_k; P\right) \end{aligned} \tag{41}$$

where the second equality follows from (12). Consequently,

$$\chi_{\mathbf{a}}(P; \mathcal{G}) = E_P\left[\left[\prod_{k=0}^{p} \pi_{a_k}\left(\overline{\mathbf{B}}_k; P\right)\right]^{-1} I_{\mathbf{a}}(\mathbf{A})\, Y\right] = E_P\left[\left[\prod_{k=0}^{p} \pi_{a_k}\left(\overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k; P\right)\right]^{-1} I_{\mathbf{a}}(\mathbf{A})\, Y\right].$$

The first equality is true because $\mathbf{B}$ is a time dependent adjustment set. The second equality, which follows from (41), proves that $(\mathbf{G}, \mathbf{B})$ is also an adjustment set.

We will now prove the following results

1.

$$\sigma^2_{\mathbf{a},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{G},\mathbf{B}}(P) = \sum_{k=0}^{p} E_P\left[\frac{I_{\overline{\mathbf{a}}_{k-1}}(\overline{\mathbf{A}}_{k-1})}{\lambda_{\overline{\mathbf{a}}_{k-1}}(\overline{\mathbf{B}}_{k-1}; P)^2}\left\{\frac{1}{\pi_{a_k}(\overline{\mathbf{B}}_k; P)} - 1\right\} var_P\left[b_{\overline{\mathbf{a}}_k}(\overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k; P) \,|\overline{\mathbf{A}}_{k-1} = \overline{\mathbf{a}}_{k-1}, \overline{\mathbf{B}}_k\right]\right].$$

2.

$$\sigma^2_{\Delta,\mathbf{B}}(P) - \sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P) = \sum_{k=0}^{p} var_P\left[t_k(\overline{\mathbf{A}}_k, \overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k, P)\right],$$

where

$$t_k(\overline{\mathbf{A}}_k, \overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k, P) \equiv \sum_{\mathbf{a}\in\mathcal{A}} c_{\mathbf{a}} \frac{I_{\overline{\mathbf{a}}_{k-1}}(\overline{\mathbf{A}}_{k-1})}{\lambda_{\overline{\mathbf{a}}_{k-1}}(\overline{\mathbf{B}}_{k-1}; P)}\left\{\frac{I_{a_k}(A_k)}{\pi_{a_k}(\overline{\mathbf{B}}_k; P)} - 1\right\}\left\{b_{\overline{\mathbf{a}}_k}(\overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k; P) - b_{\overline{\mathbf{a}}_{k-1}}(\overline{\mathbf{G}}_{k-1}, \overline{\mathbf{B}}_{k-1}; P)\right\}.$$

These imply that

$$\sigma^2_{\mathbf{a},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{G},\mathbf{B}}(P) \geq 0 \quad \text{and} \quad \sigma^2_{\Delta,\mathbf{B}}(P) - \sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P) \geq 0.$$

For $k = 0, \ldots, p$, let

$$\Lambda_k(P) \equiv \left\{q_k(\overline{\mathbf{A}}_k, \overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k) : E_P\left[q_k(\overline{\mathbf{A}}_k, \overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k) \,|\overline{\mathbf{A}}_{k-1}, \overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k\right] = 0\right\}.$$

Note that for any $0 \leq k \neq k' \leq p$, the elements of $\Lambda_k(P)$ are uncorrelated under P with those of $\Lambda_{k'}(P)$. Note also that for any function $s_k(\overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k)$ and any $P \in \mathcal{M}(\mathcal{G})$, the function

$$r_k(\overline{\mathbf{A}}_k, \overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k; s_k, P) \equiv \frac{I_{\overline{\mathbf{a}}_{k-1}}(\overline{\mathbf{A}}_{k-1})}{\lambda_{\overline{\mathbf{a}}_{k-1}}(\overline{\mathbf{B}}_{k-1}; P)}\left\{\frac{I_{a_k}(A_k)}{\pi_{a_k}(\overline{\mathbf{B}}_k; P)} - 1\right\} s_k(\overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k)$$

belongs to $\Lambda_k(P)$ because $E_P\left[\frac{I_{a_k}(A_k)}{\pi_{a_k}(\overline{\mathbf{B}}_k; P)} - 1 \middle| \overline{\mathbf{A}}_{k-1}, \overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k\right] = 0$ by (12). Next, write $\psi_{P,\mathbf{a}}(\mathbf{B}; \mathcal{G}) = \psi_{P,\mathbf{a}}(\mathbf{G}, \mathbf{B}; \mathcal{G}) + \sum_{k=0}^{p} r_k\left(\overline{\mathbf{A}}_k, \overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k; s^*_{\mathbf{a},k}, P\right)$ where $s^*_{\mathbf{a},k}(\overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k) \equiv b_{\overline{\mathbf{a}}_k}(\overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k; P) - b_{\overline{\mathbf{a}}_k}(\overline{\mathbf{B}}_k; P)$. Noting that $\psi_{P,\mathbf{a}}(\mathbf{G}, \mathbf{B}; \mathcal{G})$ is uncorrelated under P with the elements of $\Lambda_k(P)$ for all $0 \leq k \leq p$ (Robins and Rotnitzky, 1992), we conclude that

$var_P\left[\psi_{P,\mathbf{a}}\left(\mathbf{B};\mathcal{G}\right)\right] = var_P\left[\psi_{P,\mathbf{a}}\left(\mathbf{G},\mathbf{B};\mathcal{G}\right)\right] + \sum_{k=0}^{p} var_P\left[r_k\left(\overline{\mathbf{A}}_k,\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k;s^*_{\mathbf{a},k},P\right)\right]$. Finally

$$
\begin{aligned}
&var_P\left[r_k\left(\overline{\mathbf{A}}_k,\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k;s^*_{\mathbf{a},k},P\right)\right] = \\
&E_P\left[\frac{I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{B}}_{k-1};P\right)^2}\left\{\frac{I_{a_k}\left(A_k\right)}{\pi_{a_k}\left(\overline{\mathbf{B}}_k;P\right)}-1\right\}^2\left\{b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k;P\right)-b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{B}}_k;P\right)\right\}^2\right] = \\
&E_P\left[\frac{I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)\left[b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k;P\right)-E_P\left[b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k;P\right)|\overline{\mathbf{a}}_{k-1},\overline{\mathbf{B}}_k\right]\right]^2}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{B}}_{k-1};P\right)^2}\left\{\frac{1}{\pi_{a_k}\left(\overline{\mathbf{B}}_k;P\right)}-1\right\}\right] = \\
&E_P\left[\frac{I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{B}}_{k-1};P\right)^2}\left\{\frac{1}{\pi_{a_k}\left(\overline{\mathbf{B}}_k;P\right)}-1\right\}var_P\left[b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k;P\right)|\overline{\mathbf{a}}_{k-1},\overline{\mathbf{B}}_k\right]\right].
\end{aligned}
$$

Next, noticing that $\mathbf{c}^T\psi_P\left(\mathbf{B};\mathcal{G}\right) = \mathbf{c}^T\psi_P\left(\mathbf{G},\mathbf{B};\mathcal{G}\right) + \sum_{k=0}^{p} t_k\left(\overline{\mathbf{A}}_k,\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k,P\right)$ and that $t_k\left(\overline{\mathbf{A}}_k,\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k,P\right) \in \Lambda_k\left(P\right)$ we obtain $\sigma^2_{\mathbf{\Delta},\mathbf{B}}\left(P\right) = var_P\left[\mathbf{c}^T\psi_P\left(\mathbf{B};\mathcal{G}\right)\right] = \sigma^2_{\mathbf{\Delta},\mathbf{G},\mathbf{B}}\left(P\right) + \sum_{k=0}^{p} var_P\left[t_k\left(\overline{\mathbf{A}}_k,\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k,P\right)\right]$. This concludes the proof of Lemma 11. ■

Proof [Proof of Lemma 12] First we show by reverse induction in k that for all $k \in \{0,1,\ldots,p\}$ it holds that

$$
b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{B}}_k,\overline{\mathbf{G}}_k;P\right) = b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k;P\right). \tag{42}
$$

This result immediately implies that $\mathbf{G} = (\mathbf{G}_0,\mathbf{G}_1,\ldots,\mathbf{G}_p)$ is a time dependent adjustment set because, $\chi_{\mathbf{a}}\left(P;\mathcal{G}\right) \equiv E_P\left[b_{\overline{\mathbf{a}}_0}\left(\overline{\mathbf{B}}_0,\overline{\mathbf{G}}_0;P\right)\right] = E_P\left[b_{\overline{\mathbf{a}}_0}\left(\overline{\mathbf{G}}_0;P\right)\right]$, where the first equality follows from the assumption that $(\mathbf{G},\mathbf{B})$ is a time dependent adjustment set and the second follows from (42) applied to $k=0$. We show that (42) holds for $k \in \{0,1,\ldots,p\}$ by reverse induction in k. First note that $b_{\overline{\mathbf{a}}_p}\left(\overline{\mathbf{B}}_p,\overline{\mathbf{G}}_p;P\right) \equiv E_P\left[Y|\mathbf{B},\mathbf{G},\overline{\mathbf{A}}_p=\overline{\mathbf{a}}_p\right] = E_P\left[Y|\mathbf{G},\overline{\mathbf{A}}_p=\overline{\mathbf{a}}_p\right] \equiv b_{\overline{\mathbf{a}}_p}\left(\overline{\mathbf{G}}_p;P\right)$ where the second equality by (13). Then (42) holds for $k=p$. Next, assume that (42) holds for $k \in \{k^*+1,\ldots,p\}$ for some $k^* \geq 0$. We will show that it holds for $k=k^*$. This follows from

$$
\begin{aligned}
b_{\overline{\mathbf{a}}_{k^*}}\left(\overline{\mathbf{B}}_{k^*},\overline{\mathbf{G}}_{k^*};P\right) &\equiv E_P\left[b_{\overline{\mathbf{a}}_{k^*}}\left(\overline{\mathbf{B}}_{k^*+1},\overline{\mathbf{G}}_{k^*+1};P\right)|\overline{\mathbf{B}}_{k^*},\overline{\mathbf{G}}_{k^*},\overline{\mathbf{A}}_{k^*}=\overline{\mathbf{a}}_{k^*}\right] \\
&= E_P\left[b_{\overline{\mathbf{a}}_{k^*}}\left(\overline{\mathbf{G}}_{k^*+1};P\right)|\overline{\mathbf{B}}_{k^*},\overline{\mathbf{G}}_{k^*},\overline{\mathbf{A}}_{k^*}=\overline{\mathbf{a}}_{k^*}\right] \\
&= E_P\left[b_{\overline{\mathbf{a}}_{k^*}}\left(\overline{\mathbf{G}}_{k^*+1};P\right)|\overline{\mathbf{G}}_{k^*},\overline{\mathbf{A}}_{k^*}=\overline{\mathbf{a}}_{k^*}\right] \equiv b_{\overline{\mathbf{a}}_{k^*}}\left(\overline{\mathbf{G}}_{k^*};P\right),
\end{aligned}
$$

where the second equality is by the inductive hypothesis and the third is by (14) applied to $j=k^*+1$. Next we show that for any $k \in \{0,\ldots,p\}$

$$
E_P\left[\left.\frac{1}{\lambda_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k;P\right)}\right|\overline{\mathbf{G}}_k,\overline{\mathbf{A}}_k=\overline{\mathbf{a}}_k\right] = \frac{1}{\lambda_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k;P\right)}. \tag{43}
$$

To do so we write for $k \in \{1, \ldots, p\}$

$$E_P\left[\left.\frac{1}{\lambda_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k; P\right)}\right| \overline{\mathbf{G}}_k, \overline{\mathbf{a}}_k\right] \pi_{a_k}\left(\overline{\mathbf{G}}_k; P\right) =$$

$$E_P\left[\left.\frac{1}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1}, \overline{\mathbf{B}}_{k-1}; P\right)} \frac{A_k}{\pi_{a_k}\left(\overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k; P\right)}\right| \overline{\mathbf{G}}_k, \overline{\mathbf{a}}_{k-1}\right] =$$

$$E_P\left[\left.\frac{1}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1}, \overline{\mathbf{B}}_{k-1}; P\right)}\right| \overline{\mathbf{G}}_k, \overline{\mathbf{a}}_{k-1}\right] = E_P\left[\left.\frac{1}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1}, \overline{\mathbf{B}}_{k-1}; P\right)}\right| \overline{\mathbf{G}}_{k-1}, \overline{\mathbf{A}}_{k-1} = \overline{\mathbf{a}}_{k-1}\right]$$

where the last equality is by (14) applied to $j = k$. Likewise,
$E_P\left[\lambda^{-1}_{\overline{\mathbf{a}}_0}\left(\overline{\mathbf{G}}_0, \overline{\mathbf{B}}_0; P\right) \middle| \overline{\mathbf{G}}_0, \overline{\mathbf{a}}_0\right] \pi_{a_0}\left(\overline{\mathbf{G}}_0; P\right) = 1$. Then, for any $k \in \{0, \ldots, p\}$

$$E_P\left[\left.\frac{1}{\lambda_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k; P\right)}\right| \overline{\mathbf{G}}_k, \overline{\mathbf{a}}_k\right] = \frac{1}{\pi_{a_k}\left(\overline{\mathbf{G}}_k; P\right)} E_P\left[\left.\frac{1}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1}, \overline{\mathbf{B}}_{k-1}; P\right)}\right| \overline{\mathbf{G}}_{k-1}, \overline{\mathbf{a}}_{k-1}\right]$$

$$= \frac{1}{\pi_{a_{k-1}}\left(\overline{\mathbf{G}}_{k-1}; P\right)} \frac{1}{\pi_{a_k}\left(\overline{\mathbf{G}}_k; P\right)} E_P\left[\left.\frac{1}{\lambda_{\overline{\mathbf{a}}_{k-2}}\left(\overline{\mathbf{G}}_{k-2}, \overline{\mathbf{B}}_{k-2}; P\right)}\right| \overline{\mathbf{G}}_{k-2}, \overline{\mathbf{a}}_{k-2}\right]$$

$$= \cdots = \frac{1}{\pi_{a_0}\left(\overline{\mathbf{G}}_0; P\right)} \frac{1}{\pi_{a_{k-1}}\left(\overline{\mathbf{G}}_{k-1}; P\right)} \frac{1}{\pi_{a_k}\left(\overline{\mathbf{G}}_k; P\right)} \equiv \frac{1}{\lambda_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k; P\right)}$$

We will now prove the following results

1.

$$\sigma^2_{\mathbf{a},\mathbf{G},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{G}}(P) = E_P\left[var_P\left[\left.\frac{I_{\mathbf{a}}(\mathbf{A})}{\lambda_{\overline{\mathbf{a}}_p}(\mathbf{G}, \mathbf{B}; P)}\left\{Y - b_{\overline{\mathbf{a}}_p}(\mathbf{G}, \mathbf{B}; P)\right\}\right| Y, \overline{\mathbf{G}}_p, \overline{\mathbf{A}}_p\right]\right] +$$

$$\sum_{k=0}^{p} E_P\left[var_P\left[\left.\frac{I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)\left\{b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k; P\right) - b_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1}, \overline{\mathbf{B}}_{k-1}; P\right)\right\}}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1}, \overline{\mathbf{B}}_{k-1}; P\right)}\right| \overline{\mathbf{G}}_k, \overline{\mathbf{A}}_{k-1}\right]\right],$$

2.

$$\sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P) - \sigma^2_{\Delta,\mathbf{G}}(P) = E_P\left[var_P\left[\left.\sum_{\mathbf{a}\in\mathcal{A}} \frac{c_{\mathbf{a}} I_{\mathbf{a}}(\mathbf{A})\left\{Y - b_{\overline{\mathbf{a}}_p}(\mathbf{G}, \mathbf{B}; P)\right\}}{\lambda_{\overline{\mathbf{a}}_p}(\mathbf{G}, \mathbf{B}; P)}\right| Y, \overline{\mathbf{G}}_p, \overline{\mathbf{A}}_p\right]\right] +$$

$$\sum_{k=0}^{p} E_P\left[var_P\left[\left.\sum_{\mathbf{a}\in\mathcal{A}} \frac{c_{\mathbf{a}} I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)\left[b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k, \overline{\mathbf{B}}_k; P\right) - b_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1}, \overline{\mathbf{B}}_{k-1}; P\right)\right]}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1}, \overline{\mathbf{B}}_{k-1}; P\right)}\right| \overline{\mathbf{G}}_k, \overline{\mathbf{A}}_{k-1}\right]\right]$$

which imply that $\sigma^2_{\mathbf{a},\mathbf{G},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{G}}(P) \geq 0$ and $\sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P) - \sigma^2_{\Delta,\mathbf{G}}(P) \geq 0$.

We note that re-arranging terms, the influence function (11) can be re-expressed as

$$\psi_{P,\mathbf{a}}(\mathbf{Z}; P) = \frac{I_{\mathbf{a}}(\mathbf{A})}{\lambda_{\overline{\mathbf{a}}_p}(\mathbf{Z}; P)}\left\{Y - b_{\overline{\mathbf{a}}_p}(\mathbf{Z}; P)\right\} + \sum_{k=0}^{p} \frac{I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)\left\{b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{Z}}_k; P\right) - b_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{Z}}_{k-1}; P\right)\right\}}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{Z}}_{k-1}; P\right)}$$

where $b_{\overline{\mathbf{a}}_{-1}}\left(\overline{\mathbf{Z}}_{-1}; P\right) \equiv \chi_{\mathbf{a}}(P; \mathcal{G})$. Furthermore, for any $\mathbf{Z}$, the terms $\frac{I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{Z}}_{k-1}; P\right)} \times$
$\left\{b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{Z}}_k; P\right) - b_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{Z}}_{k-1}; P\right)\right\}, k \in \{0, \ldots, p\}$ and $\frac{I_{\mathbf{a}}(\mathbf{A})}{\lambda_{\overline{\mathbf{a}}_p}(\mathbf{Z}; P)}\left\{Y - b_{\overline{\mathbf{a}}_p}(\mathbf{Z}; P)\right\}$ are mutually

uncorrelated under P. Then, $var_P\left[\psi_{P,\mathbf{a}}\left(\mathbf{G},\mathbf{B};P\right)\right]=var_P\left[\frac{I_{\mathbf{a}}(\mathbf{A})}{\lambda_{\overline{\mathbf{a}}_p}(\mathbf{G},\mathbf{B};P)}\left\{Y-b_{\overline{\mathbf{a}}_p}\left(\mathbf{G},\mathbf{B};P\right)\right\}\right]+$
$\sum_{k=0}^{p}var_P\left[\frac{I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1},\overline{\mathbf{B}}_{k-1};P\right)}\left\{b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k;P\right)-b_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1},\overline{\mathbf{B}}_{k-1};P\right)\right\}\right]$. Now,
$E_P\left[\frac{I_{\mathbf{a}}(\mathbf{A})}{\lambda_{\overline{\mathbf{a}}_p}(\mathbf{G},\mathbf{B};P)}\left\{Y-b_{\overline{\mathbf{a}}_p}\left(\mathbf{G},\mathbf{B};P\right)\right\}\middle|Y,\overline{\mathbf{G}}_p,\overline{\mathbf{A}}_p\right]=I_{\mathbf{a}}\left(\mathbf{A}\right)\left\{Y-b_{\overline{\mathbf{a}}_p}\left(\mathbf{G};P\right)\right\}\times$
$E_P\left[\frac{1}{\lambda_{\overline{\mathbf{a}}_p}(\overline{\mathbf{G}}_p,\overline{\mathbf{B}}_p;P)}\middle|Y,\overline{\mathbf{G}}_p,\overline{\mathbf{a}}_p\right]=I_{\mathbf{a}}\left(\mathbf{A}\right)\left\{Y-b_{\overline{\mathbf{a}}_p}\left(\mathbf{G};P\right)\right\}E_P\left[\frac{1}{\lambda_{\overline{\mathbf{a}}_p}(\overline{\mathbf{G}}_p,\overline{\mathbf{B}}_p;P)}\middle|\overline{\mathbf{G}}_p,=\overline{\mathbf{a}}_p\right]=$
$\frac{I_{\mathbf{a}}(\mathbf{A})}{\lambda_{\overline{\mathbf{a}}_p}(\mathbf{G};P)}\left\{Y-b_{\overline{\mathbf{a}}_p}\left(\mathbf{G};P\right)\right\}$, where the first equality is by (42) applied to $k=p$, the second is by (13) and the third is by (43) applied to $k=p$. Also, for any $k\in\{0,\ldots,p\}$

$$E_P\left[\frac{I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1},\overline{\mathbf{B}}_{k-1};P\right)}\left\{b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k;P\right)-b_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1},\overline{\mathbf{B}}_{k-1};P\right)\right\}\middle|\overline{\mathbf{G}}_k,\overline{\mathbf{A}}_{k-1}\right]=$$
$$I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)\left\{b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k;P\right)-b_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1};P\right)\right\}E_P\left[\frac{1}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1},\overline{\mathbf{B}}_{k-1};P\right)}\middle|\overline{\mathbf{G}}_k,\overline{\mathbf{a}}_{k-1}\right]=$$
$$I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)\left\{b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k;P\right)-b_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1};P\right)\right\}E_P\left[\frac{1}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1},\overline{\mathbf{B}}_{k-1};P\right)}\middle|\overline{\mathbf{G}}_{k-1},\overline{\mathbf{a}}_{k-1}\right]=$$
$$\frac{I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1};P\right)}\left\{b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k;P\right)-b_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1};P\right)\right\}$$

the first equality is by (42) , the second is by (14) and the third is by (43) and where, recall, for $k=0$, $I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)\equiv\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1};P\right)\equiv 1$ and $b_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1};P\right)\equiv\chi_{\mathbf{a}}\left(P;\mathcal{G}\right)$. Then

$$\begin{aligned}
&var_P\left[\psi_{P,\mathbf{a}}\left(\mathbf{G},\mathbf{B};P\right)\right]=var_P\left[\frac{I_{\mathbf{a}}\left(\mathbf{A}\right)}{\lambda_{\overline{\mathbf{a}}_p}\left(\mathbf{G},\mathbf{B};P\right)}\left\{Y-b_{\overline{\mathbf{a}}_p}\left(\mathbf{G},\mathbf{B};P\right)\right\}\right]\\
&+\sum_{k=0}^{p}var_P\left[\frac{I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1},\overline{\mathbf{B}}_{k-1};P\right)}\left\{b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k;P\right)-b_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1},\overline{\mathbf{B}}_{k-1};P\right)\right\}\right]\\
&=var_P\left\{E_P\left[\frac{I_{\mathbf{a}}\left(\mathbf{A}\right)\left\{Y-b_{\overline{\mathbf{a}}_p}\left(\mathbf{G},\mathbf{B};P\right)\right\}}{\lambda_{\overline{\mathbf{a}}_p}\left(\mathbf{G},\mathbf{B};P\right)}\middle|Y,\overline{\mathbf{G}}_p,\overline{\mathbf{A}}_p\right]\right\}\\
&+E_P\left\{var_P\left[\frac{I_{\mathbf{a}}\left(\mathbf{A}\right)\left\{Y-b_{\overline{\mathbf{a}}_p}\left(\mathbf{G},\mathbf{B};P\right)\right\}}{\lambda_{\overline{\mathbf{a}}_p}\left(\mathbf{G},\mathbf{B};P\right)}\middle|Y,\overline{\mathbf{G}}_p,\overline{\mathbf{A}}_p\right]\right\}\\
&+\sum_{k=0}^{p}var_P\left\{E_P\left[\frac{I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)\left\{b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k;P\right)-b_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1},\overline{\mathbf{B}}_{k-1};P\right)\right\}}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1},\overline{\mathbf{B}}_{k-1};P\right)}\middle|\overline{\mathbf{G}}_k,\overline{\mathbf{A}}_{k-1}\right]\right\}\\
&+\sum_{k=0}^{p}E_P\left\{var_P\left[\frac{I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)\left\{b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k;P\right)-b_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1},\overline{\mathbf{B}}_{k-1};P\right)\right\}}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1},\overline{\mathbf{B}}_{k-1};P\right)}\middle|\overline{\mathbf{G}}_k,\overline{\mathbf{A}}_{k-1}\right]\right\}\\
&=var_P\left[\psi_{P,\mathbf{a}}\left(\mathbf{G};P\right)\right]+E_P\left\{var_P\left[\frac{I_{\mathbf{a}}\left(\mathbf{A}\right)}{\lambda_{\overline{\mathbf{a}}_p}\left(\mathbf{G},\mathbf{B};P\right)}\left\{Y-b_{\overline{\mathbf{a}}_p}\left(\mathbf{G},\mathbf{B};P\right)\right\}\middle|Y,\overline{\mathbf{G}}_p,\overline{\mathbf{A}}_p\right]\right\}\\
&+\sum_{k=0}^{p}E_P\left\{var_P\left[\frac{I_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{A}}_{k-1}\right)\left\{b_{\overline{\mathbf{a}}_k}\left(\overline{\mathbf{G}}_k,\overline{\mathbf{B}}_k;P\right)-b_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1},\overline{\mathbf{B}}_{k-1};P\right)\right\}}{\lambda_{\overline{\mathbf{a}}_{k-1}}\left(\overline{\mathbf{G}}_{k-1},\overline{\mathbf{B}}_{k-1};P\right)}\middle|\overline{\mathbf{G}}_k,\overline{\mathbf{A}}_{k-1}\right]\right\}.
\end{aligned}$$

The formula for $\sigma^2_{\mathbf{a},\mathbf{G},\mathbf{B}}(P) - \sigma^2_{\mathbf{a},\mathbf{G}}(P)$ follows by recalling that for any $\mathbf{Z}$, $\sigma^2_{\mathbf{a},\mathbf{Z},P} = var_P[\psi_{P,\mathbf{a}}(\mathbf{Z};P)]$.

We derive the formula for $\sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P) - \sigma^2_{\Delta,\mathbf{G}}(P)$ analogously. Specifically, for any $\mathbf{Z}$,

$$\sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P) = var_P\left[\sum_{\mathbf{a}\in\mathcal{A}} c_{\mathbf{a}}\psi_{P,\mathbf{a}}(\mathbf{Z};\mathcal{G})\right].$$

But,

$$\sum_{\mathbf{a}\in\mathcal{A}} c_{\mathbf{a}}\psi_{P,\mathbf{a}}(\mathbf{Z};\mathcal{G}) =$$
$$\left\{\sum_{\mathbf{a}\in\mathcal{A}}\left[c_{\mathbf{a}}\frac{I_{\mathbf{a}}(\mathbf{A})}{\lambda_{\overline{\mathbf{a}}_p}(\mathbf{Z};P)}\left\{Y - b_{\overline{\mathbf{a}}_p}(\mathbf{Z};P)\right\}\right]\right\} +$$
$$\sum_{k=0}^{p}\left\{\sum_{\mathbf{a}\in\mathcal{A}}\left[c_{\mathbf{a}}\frac{I_{\overline{\mathbf{a}}_{k-1}}(\overline{\mathbf{A}}_{k-1})}{\lambda_{\overline{\mathbf{a}}_{k-1}}(\overline{\mathbf{Z}}_{k-1};P)} \times \left\{b_{\overline{\mathbf{a}}_k}(\overline{\mathbf{Z}}_k;P) - b_{\overline{\mathbf{a}}_{k-1}}(\overline{\mathbf{Z}}_{k-1};P)\right\}\right]\right\}.$$

It is easy to check that the terms in between curly brackets in the last display are uncorrelated. Thus the formula for $\sigma^2_{\Delta,\mathbf{G},\mathbf{B}}(P) - \sigma^2_{\Delta,\mathbf{G}}(P)$ is derived verbatim as in the sequence of equalities for $var_P[\psi_{P,\mathbf{a}}(\mathbf{G};P)]$ in the display above, except that $\sum_{\mathbf{a}\in\mathcal{A}} c_{\mathbf{a}}$ is included in front of each expression between squared brackets.

This concludes the proof of Lemma 12. ■

A.2 Proofs of results in Section 6

Lemma 23 *For $\mathcal{G}$ the DAG in Figure 6, let $P_\alpha \in \mathcal{M}(\mathcal{G})$ satisfy (1) $b_a(\mathbf{O};P_\alpha) = O_1 + O_2 + \alpha O_1 O_2$, (2) $E_{P_\alpha}(O_1) = E_{P_\alpha}(O_2) = 0$, (3) $E_{P_\alpha}(O_1^2) = E_{P_\alpha}(O_2^2) = 1$, and (4) There exists a fixed $C > 0$ independent of α such that $var_{P_\alpha}(Y \mid A = a, \mathbf{O}) \leq C$ and $\pi_a(\mathbf{O}_{min};P_\alpha) \geq 1/C$. Then $var_{P_\alpha}[\psi_{P_\alpha,a}(\mathbf{O};\mathcal{G})]/var_{P_\alpha}\left[\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})\right] \underset{|\alpha|\to\infty}{\to} \infty$.*

Proof Recall that $\psi_{,a}(\mathbf{O};\mathcal{G}) \equiv \frac{I_a(A)}{\pi_a(\mathbf{O}_{min};P)}\{Y - b_a(\mathbf{O};P)\} + b_a(\mathbf{O};P) - \chi_a(P;\mathcal{G})$ is an influence function of $\chi_a(P;\mathcal{G})$ under the Bayesian Network $\mathcal{M}(\mathcal{G})$. This is because by $\mathbf{O}$ being an adjustment set we know that for all $P \in \mathcal{M}(\mathcal{G})$, $\chi_a(P;\mathcal{G}) = E_P[E_p(Y|A=a,\mathbf{O})]$. Then, $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G}) = \Pi[\psi_{P,a}(\mathbf{O};\mathcal{G})|\Lambda(P)]$ where $\Lambda(P)$ is the tangent space of model $\mathcal{M}(\mathcal{G})$ at P. Consequently, $\Delta_{P_\alpha}(\mathbf{O}) = \Pi\left[\psi_{P_\alpha,a}(\mathbf{O};\mathcal{G})|\Lambda(P_\alpha)^\perp\right]$ and by Pythagoras's Theorem, we have $var_{P_\alpha}\left[\chi^1_{P_\alpha,a,eff}(\mathbf{V};\mathcal{G})\right] = var_{P_\alpha}[\psi_{P_\alpha,a}(\mathbf{O};\mathcal{G})] - var_{P_\alpha}[\Delta_{P_\alpha}(\mathbf{O})]$. Therefore, $\frac{var_{P_\alpha}\left[\chi^1_{P_\alpha,a,eff}(\mathbf{V};\mathcal{G})\right]}{var_{P_\alpha}\left[\psi_{P_\alpha,a}(\mathbf{O};\mathcal{G})\right]} = 1 - \frac{var_{P_\alpha}[\Delta_{P_\alpha}(\mathbf{O})]}{var_{P_\alpha}\left[\psi_{P_\alpha,a}(\mathbf{O};\mathcal{G})\right]}$. Now, O_1 and O_2 are marginally independent under all $P \in \mathcal{M}(\mathcal{G})$. Since $E_{P_\alpha}(O_1) = E_{P_\alpha}(O_2) = 0$, we have that $E_{P_\alpha}[b_a(\mathbf{O};P_\alpha)|O_1] = O_1$, $E_{P_\alpha}[b_a(\mathbf{O};P_\alpha)|O_2] = O_2$ and $E_{P_\alpha}[b_a(\mathbf{O};P_\alpha)] = 0$. Thus, $\Delta_{P_\alpha}(\mathbf{O}) \equiv b_a(\mathbf{O};P_\alpha) - E_{P_\alpha}[b_a(\mathbf{O};P_\alpha)|O_1] - E_{P_\alpha}[b_a(\mathbf{O};P_\alpha)|O_2] + E_{P_\alpha}[b_a(\mathbf{O};P_\alpha)] = \alpha O_1 O_2$. Consequently, $var_{P_\alpha}[\Delta_{P_\alpha}(\mathbf{O})] = \alpha^2 E_{P_\alpha}\left[O_1^2 O_2^2\right] = \alpha^2$. On the other hand, $var_{P_\alpha}[\psi_{P,a}(\mathbf{V};\mathcal{G})] =$

$var_{P_\alpha}\left[\frac{I_a(A)}{\pi_a(\mathbf{O}_{min};P_\alpha)}\left\{Y-b_a\left(\mathbf{O};P_\alpha\right)\right\}\right]+var_{P_\alpha}\left[b_a\left(\mathbf{O};P_\alpha\right)-\chi_a(P_\alpha;\mathcal{G})\right]=$
$E_{P_\alpha}\left[\frac{I_a(A)}{\pi_a^2(\mathbf{O}_{min};P_\alpha)}var_{P_\alpha}(Y\mid A=a,\mathbf{O})\right]+E_{P_\alpha}\left[b_a^2\left(\mathbf{O};P_\alpha\right)\right]=$
$E_{P_\alpha}\left[\frac{I_a(A)}{\pi_a^2(\mathbf{O}_{min};P_\alpha)}var_{P_\alpha}(Y\mid A=a,\mathbf{O})\right]+2+\alpha^2$. where the last equality follows because $E_{P_\alpha}\left[b_a^2\left(\mathbf{O};P_\alpha\right)\right]=E_{P_\alpha}\left[\left\{O_1+O_2+\alpha O_1O_2\right\}^2\right]=2+\alpha^2$ since O_1 and O_2 have zero mean, unit variance, and are uncorrelated under P_α. Since by assumption $var_{P_\alpha}(Y\mid A=a,\mathbf{O})\leq C$ and $\pi_a(\mathbf{O}_{min};P_\alpha)\geq 1/C$, we have $E_{P_\alpha}\left[\frac{I_a(A)}{\pi_a^2(\mathbf{O}_{min};P_\alpha)}var_{P_\alpha}(Y\mid A=a,\mathbf{O})\right]\leq C^3$. Consequently,

$$\frac{var_{P_\alpha}\left[\Delta_{P_\alpha}\left(\mathbf{O}\right)\right]}{var_{P_\alpha}\left[\psi_{P_\alpha,a}(\mathbf{O};\mathcal{G})\right]}=\frac{\alpha^2}{E_{P_\alpha}\left[I_a(A)\pi_a^{-2}(\mathbf{O}_{min};P_\alpha)var_{P_\alpha}(Y\mid A=a,\mathbf{O})\right]+2+\alpha^2}\underset{|\alpha|\to\infty}{\to}1$$

and therefore $\frac{var_{P_\alpha}\left[\chi^1_{P_\alpha,a,eff}(\mathbf{V};\mathcal{G})\right]}{var_{P_\alpha}\left[\psi_{P_\alpha,a}(\mathbf{O};\mathcal{G})\right]}=1-\frac{var_{P_\alpha}\left[\Delta_{P_\alpha}(\mathbf{O})\right]}{var_{P_\alpha}\left[\psi_{P_\alpha,a}(\mathbf{O};\mathcal{G})\right]}\underset{|\alpha|\to\infty}{\to}0.$ ■

Lemma 24 *Let $\mathcal{G}$ be a DAG with vertex set that stands for a random vector $\mathbf{V}=(V_1,...,V_s)$. Suppose that the laws in the Bayesian Network $\mathcal{M}(\mathcal{G})$ are dominated by some measure μ. Then the tangent space of model $\mathcal{M}(\mathcal{G})$ at a law P is given by $\Lambda(P)\equiv\oplus_{j=1}^s\Lambda_j(P)$ where*

$$\Lambda_j(P)=\left\{G\equiv g\left(V_j,\mathrm{pa}_{\mathcal{G}}\left(V_j\right)\right)\in L_2(P):E_P\left[G|\,\mathrm{pa}_{\mathcal{G}}\left(V_j\right)\right]=0\right\}. \tag{44}$$

Proof For any $P\in\mathcal{M}(\mathcal{G})$ let p denote, any version of, the density of P with respect to μ. For any $P\in\mathcal{M}(\mathcal{G})$, $p(\mathbf{V})$ factors as $p(\mathbf{V})=\prod_{k=1}^s p_k\left(V_k|\mathrm{pa}_{\mathcal{G}}\left(V_k\right)\right)$ where p_j is, any version of, the conditional density of V_j given $\mathrm{pa}_{\mathcal{G}}\left(V_j\right)$. Lemma 1.6 of Van der Laan and Robins (2003), implies that the tangent space of model $\mathcal{M}(\mathcal{G})$ at a law P is given by $\Lambda\equiv\oplus_{j=1}^s\Lambda_j$ where Λ_j is the closed linear span of scores of one dimensional regular parametric submodels $t\to p(\mathbf{V};t)=p_j\left(V_j|\mathrm{pa}_{\mathcal{G}}\left(V_j\right);t\right)\prod_{k=1,k\neq j}^s p_k\left(V_k|\mathrm{pa}_{\mathcal{G}}\left(V_k\right)\right)$. Such Λ_j is equal to the set in the right hand side of (44) because model $\mathcal{M}(\mathcal{G})$ does not impose restrictions on the law $p_j\left(V_j|\mathrm{pa}_{\mathcal{G}}\left(V_j\right)\right)$ (Tsiatis (2007), Theorem 4.5). This concludes the proof. ■

In the next proofs we will use the following definitions.

Definition 25

$\mathbf{F}(A,Y,\mathcal{G})\equiv\{V_j\in\mathbf{V}:\exists$ *a path between A and Y in $\mathcal{G}$ that has V_j as its only fork*$\}$,
$\mathrm{dir}(A,Y,\mathcal{G})\equiv\{Y\}\cup$
$\{V_j\in\mathbf{V}:V_j$ *has a directed path to Y in $\mathcal{G}$ that does not intersect A*$\}\setminus\mathbf{F}(A,Y,\mathcal{G})$.

Proof [Proof of Theorem 14] Let $R_{P,a,\mathcal{G}}\equiv\left\{\frac{I_a(A)}{\pi_a\left(\mathrm{pa}_{\mathcal{G}}(A);P\right)}-1\right\}b_a\left(\mathrm{pa}_{\mathcal{G}}\left(A\right);P\right)$. Then, $\psi_{P,a}\left[\mathrm{pa}_{\mathcal{G}}(A);\mathcal{G}\right]=J_{P,a,\mathcal{G}}-R_{P,a,\mathcal{G}}-\chi_a(P;\mathcal{G})$ is an influence function for $\chi_a(P;\mathcal{G})$ in model $\mathcal{M}(\mathcal{G})$ because it is the unique influence function for $\chi_a(P;\mathcal{G})$ in the non-parametric model

that does not impose any restrictions on P. Consequently, with $\Lambda(P)$ denoting the tangent set of model $\mathcal{M}(\mathcal{G})$ at $P \in \mathcal{M}(\mathcal{G})$, we have by Lemma 24 that $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G}) = \Pi\left[\psi_{P,a}\left[\mathrm{pa}_{\mathcal{G}}(A);\mathcal{G}\right]|\Lambda(P)\right] = \sum_{j=1}^{s}\Pi\left[\psi_{P,a}\left[\mathrm{pa}_{\mathcal{G}}(A);\mathcal{G}\right]|\Lambda_j(P)\right]$, where $\Lambda_j(P)$ is defined (44). The identity (18) is proved if we show that (i) $\Pi\left[\psi_{P,a}\left[\mathrm{pa}_{\mathcal{G}}(A);\mathcal{G}\right]|\Lambda_j(P)\right] = 0$ for $V_j = A$, (ii) $\Pi\left[\psi_{P,a}\left[\mathrm{pa}_{\mathcal{G}}(A);\mathcal{G}\right]|\Lambda_j(P)\right] = E_P\left[J_{P,a,\mathcal{G}} \mid V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] - E_P\left[J_{P,a,\mathcal{G}} \mid \mathrm{pa}_{\mathcal{G}}(V_j)\right]$ for any $V_j \neq A$ and (iii) $E_P\left[J_{P,a,\mathcal{G}}|V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] - E_P\left[J_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(V_j)\right] = 0$ for any $V_j \in \mathrm{indir}(A,Y,\mathcal{G}) \cup \mathrm{an}^c_{\mathcal{G}}(\{A,Y\})$.

Assertion (i) holds because since $\chi_a(P;\mathcal{G})$ does not depend on the law of A given $\mathrm{pa}_{\mathcal{G}}(A)$, then $\psi_{P,a}\left[\mathrm{pa}_{\mathcal{G}}(A);\mathcal{G}\right]$ is orthogonal to the scores for all regular parametric submodels for the law A given $\mathrm{pa}_{\mathcal{G}}(A)$.

To show assertion (ii), first notice that for any random variable $U, \Pi[U|\Lambda_j(P)] = E_P\left[U \mid V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] - E_P\left[U \mid \mathrm{pa}_{\mathcal{G}}(V_j)\right]$. Then, assertion (ii) is proved if we prove that for any $V_j \neq A, E_P\left[R_{P,a,\mathcal{G}}|\, V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] - E_P\left[R_{P,a\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(V_j)\right] = 0$. If $V_j \in \mathrm{de}_{\mathcal{G}}(A)$ this follows from the fact that $\left[A, \mathrm{pa}_{\mathcal{G}}(A)\right] \perp\!\!\!\perp V_j|\mathrm{pa}_{\mathcal{G}}(V_j)$ by the Local Markov property and the fact that $R_{P,a,\mathcal{G}}$ depends only on A and $\mathrm{pa}_{\mathcal{G}}(A)$. On the other hand, if $V_j \in \mathrm{de}^c_{\mathcal{G}}(A)$, the result follows from $E_P\left[R_{P,a,\mathcal{G}}|\, V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] =$
$E_P\left[\left\{\frac{E_P\left[I_a(A)|\mathrm{pa}_{\mathcal{G}}(A),V_j,\mathrm{pa}_{\mathcal{G}}(V_j)\right]}{\pi_a\left(\mathrm{pa}_{\mathcal{G}}(A);P\right)} - 1\right\} b_a\left(\mathrm{pa}_{\mathcal{G}}(A);P\right)\middle| V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] = 0$. where the second equality holds because $E_P\left[I_a(A)|\mathrm{pa}_{\mathcal{G}}(A)\right] = E_P\left[I_a(A)|\mathrm{pa}_{\mathcal{G}}(A), V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right]$ since $A \perp\!\!\!\perp \left[\left[V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] \backslash \mathrm{pa}_{\mathcal{G}}(A)\right] |\mathrm{pa}_{\mathcal{G}}(A)$ by the Local Markov property.

Turn now to the proof of assertion (iii). Let $V_j \in \mathrm{indir}(A,Y,\mathcal{G})$. We will show that $E_P\left[J_{P,\mathcal{G}} \mid V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right]$ does not depend on V_j. Let $\mathbf{F} \equiv \mathbf{F}(A,Y,\mathcal{G})$. We begin by noting the following: $\mathbf{F} \cup \{V_j\} \cup \mathrm{pa}_{\mathcal{G}}(V_j)$ is comprised of non-descendants of A. This is because V_j is a non-descendant of A by assumption, since A is a descendant of V_j. This implies that $\mathrm{pa}_{\mathcal{G}}(V_j)$ is a non-descendant of A. Also, any node in $\mathbf{F}$ is, by definition, an ancestor of a parent of A, therefore it cannot be a descendant of A. Then, by the Local Markov property, $E_P\left[I_a(A)|\,\mathrm{pa}_{\mathcal{G}}(A), \mathbf{F}, V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] = E_P\left[I_a(A) \mid \mathrm{pa}_{\mathcal{G}}(A)\right] = \pi\left(\mathrm{pa}_{\mathcal{G}}(A);P\right)$. Thus, $E_P\left[J_{P,\mathcal{G}} \mid V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] = E_P\left[E_P\left[Y|A=a, \mathrm{pa}_{\mathcal{G}}(A), \mathbf{F}, V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] \mid V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right]$. We will show next that $E_P\left[Y|A=a, \mathrm{pa}_{\mathcal{G}}(A), \mathbf{F}, V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] = E_P\left[Y|A=a, \mathbf{F}\right]$. To do so, it suffices to show that

$$Y \perp\!\!\!\perp_{\mathcal{G}} \left[\{V_j\} \cup \mathrm{pa}_{\mathcal{G}}(V_j) \cup \mathrm{pa}_{\mathcal{G}}(A)\right] \backslash \mathbf{F} \big| A, \mathbf{F}. \tag{45}$$

Note that $\left[\{V_j\} \cup \mathrm{pa}_{\mathcal{G}}(V_j) \cup \mathrm{pa}_{\mathcal{G}}(A)\right] \backslash \mathbf{F} \subset \mathrm{indir}(A,Y,\mathcal{G})$. Then by Lemma 29 equation (45) holds. Hence $E_P\left[J_{P,\mathcal{G}} \mid V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] = E_P\left[E_P\left[Y|A=a, \mathbf{F}\right] \mid V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right]$. Now note that vertices in $\mathbf{F}$ cannot be descendants of V_j since, if $V \in \mathbf{F}$ were a descendant of V_j, then there would be a directed path from V_j to Y that does not intersect A, a contradiction. Hence by the Local Markov Property $V_j \perp\!\!\!\perp \mathbf{F} \mid \mathrm{pa}_{\mathcal{G}}(V_j)$.Thus $E_P\left[J_{P,\mathcal{G}} \mid V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] = E_P\left[E_P\left[Y|A=a, \mathbf{F}\right] \mid V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] = E_P\left[E_P\left[Y|A=a, \mathbf{F}\right] \mid \mathrm{pa}_{\mathcal{G}}(V_j)\right]$ which does not depend on V_j.

Next, let $V_j \in \mathrm{an}^c_{\mathcal{G}}(\{A,Y\})$. Then $\mathrm{pa}_{\mathcal{G}}(A), A, Y$ are non-descendants of V_j and thus by the Local Markov Property $V_j \perp\!\!\!\perp \mathrm{pa}_{\mathcal{G}}(A), A, Y \mid \mathrm{pa}_{\mathcal{G}}(V_j)$. Therefore, since $J_{P,a,\mathcal{G}}$ is a function of only $\mathrm{pa}_{\mathcal{G}}(A), A, Y, E_P\left[J_{P,a,\mathcal{G}}|V_j, \mathrm{pa}_{\mathcal{G}}(V_j)\right] - E_P\left[J_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(V_j)\right] = 0$. This concludes the proof of assertion (iii) and, consequently, of the identity (18).

Turn now to the proof that $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$ does not depend on any $V \in \text{irrel}(A,Y,\mathcal{G})$. Take $V \in \text{irrel}(A,Y,\mathcal{G})$ and $W \in \text{ch}_{\mathcal{G}}(V) \setminus \{A\}$. We will show next that $W \in \text{irrel}(A,Y,\mathcal{G})$. This, together with the last display, will imply that $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$ is a function only of $\mathbf{V}_{marg} = \mathbf{V} \setminus \left\{\text{an}^c_{\mathcal{G}}(\{A,Y\}) \cup \text{indir}(A,Y,\mathcal{G})\right\}$. This is because the only way in which $V \in \text{irrel}(A,Y,\mathcal{G})$ can appear in $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$ is if it belongs to the parent set of a node W that is not in $\text{irrel}(A,Y,\mathcal{G}) \cup \{A\}$.

Assume first that $W \in \text{an}_{\mathcal{G}}(A) \setminus \{A\}$. Then $W \in \text{indir}(A,Y,\mathcal{G})$, since W is a child of V and $V \in \text{irrel}(A,Y,\mathcal{G})$. Assume next that $W \notin \text{an}_{\mathcal{G}}(A)$. We claim that this implies that $W \notin \text{an}_{\mathcal{G}}(Y)$. Indeed, if $W \in \text{an}_{\mathcal{G}}(Y)$, there exists a directed path from W to Y that does not intersect A. Since W is a child of V, this implies that $V \notin \text{irrel}(A,Y,\mathcal{G})$, contradicting the assumption that $V \in \text{irrel}(A,Y,\mathcal{G})$. Thus, $W \notin \text{an}_{\mathcal{G}}(A)$ implies $W \notin \text{an}_{\mathcal{G}}(Y)$. Hence, if $W \notin \text{an}_{\mathcal{G}}(A)$ then $W \in \text{irrel}(A,Y,\mathcal{G})$. Consequently, in all cases $W \in \text{irrel}(A,Y,\mathcal{G})$, which is what we wanted to show. ■

Proof [Proof of Proposition 17] Let $P' \in \mathcal{M}'$ and $P \in \mathcal{M}$ with marginal law P'. Let $\mathbf{V}^c = \mathbf{V} \setminus \mathbf{V}'$. Let $t \in [0,\varepsilon] \to P_t$ be a regular parametric submodel of $\mathcal{M}$ with $P_{t=0} = P$ and score S. Decompose S as $S_{\mathbf{V}'} + S_{\mathbf{V}^c|\mathbf{V}'}$ where $S_{\mathbf{V}'}$ is the score in the induced regular parametric submodel $t \in (0,\varepsilon] \to P'_t$ of $\mathcal{M}'$ with $P'_{t=0} = P'$. Then $\frac{d}{dt}\chi(P_t)\big|_{t=0} = E_P\left[\chi^1_{P,eff}S\right] = E_P\left[\chi^1_{P,eff}S_{\mathbf{V}'}\right] + E_P\left[\chi^1_{P,eff}S_{\mathbf{V}^c|\mathbf{V}'}\right] = E_P\left[\chi^1_{P,eff}S_{\mathbf{V}'}\right]$, where the last equality follows because $S_{\mathbf{V}^c|\mathbf{V}'}$ is a conditional score for the law of $\mathbf{V}^c|\mathbf{V}'$ and, by assumption, $\chi^1_{P,eff}$ is a function of $\mathbf{V}'$ only. On the other hand, $\frac{d}{dt}\chi(P_t)\big|_{t=0} = \frac{d}{dt}\nu(P'_t)\big|_{t=0}$ because by assumption, $\chi(P_t) = \nu(P'_t)$. Then, $\chi^1_{P,eff}$ is an influence function for $\nu(P')$. Now let Λ' be the tangent space for model $\mathcal{M}'$ at P'. Then, $\Lambda = \Lambda' \oplus$ the closed linear span of $\left\{S_{\mathbf{V}^c|\mathbf{V}'} : S_{\mathbf{V}^c|\mathbf{V}'} \text{ is a conditional score under model } \mathcal{M}\right\}$. Since $E_P\left[\chi^1_{P,eff}S_{\mathbf{V}^c|\mathbf{V}'}\right] = 0$ for all conditional scores $S_{\mathbf{V}^c|\mathbf{V}'}$ we conclude that $\chi^1_{P,eff}$ is in Λ' and consequently, it is the efficient influence function $\nu^1_{P',eff}$. ■

Proof [Proof of Lemma 16] We will use the following property which can be shown straightforwardly. Let $\mathcal{G}^1$, $\mathcal{G}^2$ and $\mathcal{G}^3$ be DAGs with vertex sets $\mathbf{V}^1, \mathbf{V}^2$ and $\mathbf{V}^3$ such that $\mathbf{V}^1 \supset \mathbf{V}^2 \supset \mathbf{V}^3$. Then,

$$\mathcal{M}\left(\mathcal{G}^1,\mathbf{V}^2\right) = \mathcal{M}\left(\mathcal{G}^2\right) \text{ and } \mathcal{M}\left(\mathcal{G}^2,\mathbf{V}^3\right) = \mathcal{M}\left(\mathcal{G}^3\right) \Rightarrow \mathcal{M}\left(\mathcal{G}^1,\mathbf{V}^3\right) = \mathcal{M}\left(\mathcal{G}^3\right) \tag{46}$$

The set $\mathbf{V} \setminus \text{an}^c_{\mathcal{G}}(\{A,Y\})$ is an ancestral set, that is, it contains all its own ancestors :$\mathbf{V} \setminus \text{an}^c_{\mathcal{G}}(\{A,Y\}) = \text{an}_{\mathcal{G}}\left(\mathbf{V} \setminus \text{an}^c_{\mathcal{G}}(\{A,Y\})\right)$. Then, by Proposition 1 (a) of Evans (2016)

$$\mathcal{M}\left(\mathcal{G},\mathbf{V} \setminus \text{an}^c_{\mathcal{G}}(\{A,Y\})\right) = \mathcal{M}\left(\mathcal{G}_{\mathbf{V} \setminus \text{an}^c_{\mathcal{G}}(\{A,Y\})}\right). \tag{47}$$

Now, let $\widetilde{\mathcal{G}}^{l+1} \equiv \mathcal{G}_{\mathbf{V} \setminus \text{an}^c_{\mathcal{G}}(\{A,Y\})}$ and let $(I_1,\ldots,I_l)$ be the set of nodes in $\text{indir}(A,Y,\widetilde{\mathcal{G}}^{l+1})$, topologically sorted with respect to $\widetilde{\mathcal{G}}^{l+1}$. Recursively define for $j = l, l-1,\ldots,1$, $\widetilde{\mathcal{G}}^j \equiv \tau\left(\widetilde{\mathcal{G}}^{j+1}, I_j\right)$. Noticing that in $\widetilde{\mathcal{G}}^{j+1}$, I_j has a sole child equal to A, then combining Lemma 1 and Lemma 3 of Evans (2016), yields that for $j = l, l-1,\ldots,1$,

$$\mathcal{M}\left(\widetilde{\mathcal{G}}^{j+1}, \mathbf{V} \setminus \left\{\text{an}^c_{\mathcal{G}}(\{A,Y\}) \cup \left(\cup_{i=j}^{l} I_i\right)\right\}\right) = \mathcal{M}\left(\widetilde{\mathcal{G}}^{j}\right). \tag{48}$$

Repeatedly invoking (46) to the equalities (47) and (48) yields $\mathcal{M}\left(\mathcal{G}, \mathbf{V} \setminus \{\mathrm{an}^c_{\mathcal{G}}(\{A,Y\}) \cup \mathrm{indir}(A,Y,\mathcal{G})\}\right) = \mathcal{M}\left(\widetilde{\mathcal{G}}^1\right)$. Since $\mathcal{G}' = \widetilde{\mathcal{G}}^1$ is the output of Algorithm 1, this finishes the proof of the first part of the Lemma.

Now note that the pruning algorithm prunes neither A nor Y. Furthermore, it neither adds new causal paths nor deletes causal paths between A and Y. Then, $\mathrm{cn}(A,Y,\mathcal{G}) = \mathrm{cn}(A,Y,\mathcal{G}')$. Also, the pruning algorithm neither adds nor deletes any vertex that is both a non-descendant of A in $\mathcal{G}$ and parent of a vertex in $\mathrm{cn}(A,Y,\mathcal{G})$ in $\mathcal{G}$. But the set of such vertices is precisely the set $\mathbf{O}(A,Y,\mathcal{G})$. This shows that $\mathbf{O}(A,Y,\mathcal{G}) = \mathbf{O}(A,Y,\mathcal{G}')$. Then, if $P \in \mathcal{M}(\mathcal{G})$, $b_a(\mathbf{O}(A,Y,\mathcal{G});P) = b_a(\mathbf{O}(A,Y,\mathcal{G}');P_{marg})$ and $\pi_a(\mathbf{O}(A,Y,\mathcal{G});P) = \pi_a(\mathbf{O}(A,Y,\mathcal{G}');P_{marg})$. Consequently, $\psi_{P,a}[\mathbf{O}(A,Y,\mathcal{G});\mathcal{G}] = \psi_{P_{marg},a}[\mathbf{O}(A,Y,\mathcal{G}');\mathcal{G}']$. But since $\mathbf{O}(A,Y,\mathcal{G})$ is an adjustment set relative to A and Y in $\mathcal{G}$ (and $\mathcal{G}'$) we have that $\chi_a(P;\mathcal{G}) = E_P[b_a(\mathbf{O}(A,Y,\mathcal{G});P)]$ and $\chi_a(P_{marg};\mathcal{G}') = E_{P_{marg}}[b_a(\mathbf{O}(A,Y,\mathcal{G}');P_{marg})]$ and thus conclude that $\chi_a(P;\mathcal{G}) = \chi_a(P_{marg};\mathcal{G}')$.

We turn next to the proof of $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G}) = \chi^1_{P_{marg},eff}(\mathbf{V}_{marg};\mathcal{G}')$. By Theorem 14, $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$ is a function only of $\mathbf{V}_{marg} = \mathbf{V} \setminus \left\{\mathrm{an}^c_{\mathcal{G}}(\{A,Y\}) \cup \mathrm{indir}(A,Y,\mathcal{G})\right\}$. Since we have already shown that $\mathcal{M}(\mathcal{G},\ \mathbf{V}_{marg}) = \mathcal{M}(\mathcal{G}')$, that $\chi_a(P_{marg};\mathcal{G}') = \chi_a(P;\mathcal{G})$, Proposition 17 implies that $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G}) = \chi^1_{P_{marg},a,eff}(\mathbf{V}_{marg};\mathcal{G}')$. ■

Lemma 26 *Let $\mathcal{M}(\mathcal{G})$ be the Bayesian Network represented by DAG $\mathcal{G}$ with vertex set $\mathbf{V}$. Assume Y and A are single disjoint vertices. Assume* $\mathrm{irrel}(A,Y,\mathcal{G}) = \emptyset$. *Then*

1. *If $J \geq 1$ then for all $j \in \{1,\ldots,J\}$*

$$E_P\left[J_{P,a,\mathcal{G}} | W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right] = E_P\left[b_a(\mathbf{O};P) | W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right].$$

2. *If $K \geq 1$ then for all $k \in \{1,\ldots,K\}$*

$$E_P\left[J_{P,a,\mathcal{G}} | M_k, \mathrm{pa}_{\mathcal{G}}(M_k)\right] = E_P\left[T_{P,a,\mathcal{G}} | M_k, \mathrm{pa}_{\mathcal{G}}(M_k)\right].$$

3.

$$E_P\left[J_{P,a,\mathcal{G}} | Y, \mathrm{pa}_{\mathcal{G}}(Y)\right] = E_P\left[T_{P,a,\mathcal{G}} | Y, \mathrm{pa}_{\mathcal{G}}(Y)\right].$$

Proof [Proof of Lemma 26] We begin with the proof of part 1). $E_P\left[\frac{I_a(A)Y}{\pi_a(\mathrm{pa}_{\mathcal{G}}(A);P)} \mid W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right]$
$= E_P\left[\frac{I_a(A)E_P\left[Y|A=a,W_j,\mathrm{pa}_{\mathcal{G}}(W_j),\mathbf{O},\mathrm{pa}_{\mathcal{G}}(A)\right]}{\pi_a(\mathrm{pa}_{\mathcal{G}}(A);P)} \mid W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right] = E_P\left[\frac{I_a(A)E_P[Y|A=a,\mathbf{O}]}{\pi_a(\mathrm{pa}_{\mathcal{G}}(A);P)} \mid W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right]$
$= E_P\left[b_a(\mathbf{O};P)\frac{E_P\left[I_a(A)|\mathbf{O},W_j,\mathrm{pa}_{\mathcal{G}}(W_j),\mathrm{pa}_{\mathcal{G}}(A))\right]}{\pi_a(\mathrm{pa}_{\mathcal{G}}(A);P)} \mid W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right] = E_P\left[b_a(\mathbf{O};P) \mid W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right],$
where the second equality holds because $Y \perp\!\!\!\perp_{\mathcal{G}} \left[\{W_j\} \cup \mathrm{pa}_{\mathcal{G}}(W_j) \cup \mathrm{pa}_{\mathcal{G}}(A)\right] \setminus \mathbf{O} \mid \mathbf{O}, A$ and the third equality holds because the set $\left[\{W_j\} \cup \mathrm{pa}_{\mathcal{G}}(W_j) \cup \mathbf{O}\right]$ is comprised of non-descendants of A and hence, by the Local Markov Property, $A \perp\!\!\!\perp_{\mathcal{G}} \left[\{W_j\} \cup \mathrm{pa}_{\mathcal{G}}(W_j) \cup \mathbf{O}\right] \setminus \mathrm{pa}_{\mathcal{G}}(A) \mid \mathrm{pa}_{\mathcal{G}}(A)$.

Next, we prove parts 2) and 3). Define $M_{K+1} = Y$. Then, for all $k = 1, ..., K+1$, $E_P\left[J_{P,a,\mathcal{G}} | M_k, \mathrm{pa}_{\mathcal{G}}(M_k)\right] =$

$E_P\left[I_a(A)Y\ E_P\left[\frac{1}{\pi_a(\mathrm{pa}_{\mathcal{G}}(A);P)}\middle| A=a,\mathbf{O},Y,M_k,\mathrm{pa}_{\mathcal{G}}(M_k)\right]\middle| M_k,\mathrm{pa}_{\mathcal{G}}(M_k)\right]=$
$E_P\left[I_a(A)Y\ E_P\left[\frac{1}{\pi_a(\mathrm{pa}_{\mathcal{G}}(A);P)}\middle| A=a,\mathbf{O}\right]\middle| M_k,\mathrm{pa}_{\mathcal{G}}(M_k)\right]=E_P\left[T_{P,a,\mathcal{G}}|\,M_k,\mathrm{pa}_{\mathcal{G}}(M_k)\right]$ where the second equality follows from $(Y,\mathbf{M})\perp\!\!\!\perp_{\mathcal{G}}\mathrm{pa}_{\mathcal{G}}(A)\setminus\mathbf{O}\mid[\mathbf{O}\cup\{A\}]$ and the fact that for any k, $\mathrm{pa}_{\mathcal{G}}(M_k)\subset\mathbf{M}\cup\{A\}\cup\mathbf{O}$, and the third equality follows because $E_P\left[\frac{1}{\pi_a(\mathrm{pa}_{\mathcal{G}}(A);P)}\middle| A=a,\mathbf{O}\right]=\frac{1}{\pi_a(\mathbf{O}_{min};P)}$ which is a consequence of Lemma 27 in Section A.4 and the definition of $\mathbf{O}_{min}$. This concludes the proof of the theorem. ■

Proof [Proof of Theorem 18] Because $\mathrm{irrel}(A,Y,\mathcal{G})=\emptyset$, we can partition the node set $\mathbf{V}$ of $\mathcal{G}$ as $\mathbf{M}\cup\mathbf{W}\cup\{A,Y\}$ where the vertices in $\mathbf{M}$ intersect at least one causal path between A and Y, that is, $\mathbf{M}$ is the set of mediators in the causal pathways between A and Y, and $\mathbf{W}$ are non-descendants of A. We can therefore sort topologically $\mathbf{V}$ as $(W_1,\ldots,W_J,A,M_1,\ldots,M_K,Y)$, where the set $\mathbf{W}=\emptyset$ if $J=0$ and the set $\mathbf{K}=\emptyset$ if $K=0$. By Theorem 14 we have

$$\begin{aligned}\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})=E_P\left[J_{P,a,\mathcal{G}}|Y,\mathrm{pa}_{\mathcal{G}}(Y)\right]-E_P\left[J_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(Y)\right]+\\ \sum_{k=1}^{K}\left\{E_P\left[J_{P,a,\mathcal{G}}|M_k,\mathrm{pa}_{\mathcal{G}}(M_k)\right]-E_P\left[J_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(M_k)\right]\right\}+\\ \sum_{j=1}^{J}\left\{E_P\left[J_{P,a,\mathcal{G}}|W_j,\mathrm{pa}_{\mathcal{G}}(W_j)\right]-E_P\left[J_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(W_j)\right]\right\},\end{aligned}$$

where we make the conventions that $\sum_{k=1}^{0}\cdot\equiv 0$, $\sum_{j=1}^{0}\cdot\equiv 0$. The theorem is proved by invoking Lemma 26 to make the replacements $E_P\left[T_{P,a,\mathcal{G}}|Y,\mathrm{pa}_{\mathcal{G}}(Y)\right]-E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(Y)\right]$ for $E_P\left[J_{P,a,\mathcal{G}}|Y,\mathrm{pa}_{\mathcal{G}}(Y)\right]-$
$E_P\left[J_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(Y)\right]$, $E_P\left[T_{P,a,\mathcal{G}}|M_k,\mathrm{pa}_{\mathcal{G}}(M_k)\right]-E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(M_k)\right]$ for $E_P\left[J_{P,a,\mathcal{G}}|M_k,\mathrm{pa}_{\mathcal{G}}(M_k)\right]-E_P\left[J_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(M_k)\right]$ and $E_P\left[b_a(\mathbf{O};P)|W_j,\mathrm{pa}_{\mathcal{G}}(W_j)\right]-$
$E_P\left[b_a(\mathbf{O};P)|\,\mathrm{pa}_{\mathcal{G}}(W_j)\right]$ for $E_P\left[J_{P,a,\mathcal{G}}|W_j,\mathrm{pa}_{\mathcal{G}}(W_j)\right]-E_P\left[J_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(W_j)\right]$. This concludes the proof of the theorem. ■

Proof [Proof of Theorem 19]

Assertion (20) is immediate when $J=K=0$, since $\mathbf{O}=\emptyset$, $b_a(\mathbf{O};P)=\chi_a(P;\mathcal{G})$, $\pi_a(\mathbf{O};P)=P(A=a)$, and consequently

$$\psi_{P,a}[\mathbf{O};\mathcal{G}]=\frac{I_a(A)}{P(A=a)}\left(Y-\chi_a(P;\mathcal{G})\right)$$

which coincides with $\chi^1_{P,a,eff}$.

Henceforth assume $J\geq 1$ or $K\geq 1$. If $J=0$, let $\chi^{1,non-desc}_{P,a,eff}(\mathbf{V};\mathcal{G})\equiv 0$. For $J\geq 1$, let

$$\begin{aligned}\chi^{1,non-desc}_{P,a,eff}(\mathbf{V};\mathcal{G})\equiv\sum_{j=2}^{J}\left\{E_P\left[b_a(\mathbf{O};P)|W_j,\mathrm{pa}_{\mathcal{G}}(W_j)\right]-E_P\left[b_a(\mathbf{O};P)|\,\mathrm{pa}_{\mathcal{G}}(W_j)\right]\right\}\\ +E_P\left[b_a(\mathbf{O};P)|W_1,\mathrm{pa}_{\mathcal{G}}(W_1)\right]-\chi_a(P;\mathcal{G}),\end{aligned}$$

where $\sum_{j=2}^{1} \cdot \equiv 0$. Furthermore, let

$$\chi_{P,a,eff}^{1,desc}(\mathbf{V};\mathcal{G}) \equiv E_P\left[T_{P,a,\mathcal{G}}|Y,\mathrm{pa}_{\mathcal{G}}(Y)\right] - E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(Y)\right] + \sum_{k=1}^{K}\left\{E_P\left[T_{P,a,\mathcal{G}}|M_k,\mathrm{pa}_{\mathcal{G}}(M_k)\right] - E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(M_k)\right]\right\},$$

where $\sum_{k=1}^{0} \cdot \equiv 0$. By Theorem 18, $\chi_{P,a,eff}^{1}(\mathbf{V};\mathcal{G}) = \chi_{P,a,eff}^{1,non-desc}(\mathbf{V};\mathcal{G}) + \chi_{P,a,eff}^{1,desc}(\mathbf{V};\mathcal{G})$.

First note that $W_J = O_T$. This holds because, since $\mathrm{irrel}(A,Y,\mathcal{G}) = \emptyset$, there exists a directed path from W_J to Y that does not intersect A. Let Q be a child of W_J in that path. Then Q cannot be in the set $\{W_1,\ldots,W_J\}$ because W_J is the last element in the topolocally ordered sequence $W_1,\ldots,W_J$ of non-descendants of A. Then $Q \in \mathbf{M}\cup\{Y\}$ which implies that $W_J \in \mathbf{O}$ and, since $(O_1,\ldots,O_T)$ is ordered topologically, we conclude that $W_J = O_T$.

Suppose now that $J = 1$. Then $O_T = O_1 = W_1$ and $\mathbf{O} = \{O_1\}$. Consequently, $E_P\left[b_a(\mathbf{O};P) \mid \mathrm{pa}_{\mathcal{G}}(W_1), W_1\right] = b_a(\mathbf{O};P)$. Thus

$$\chi_{P,a,eff}^{1,non-desc}(\mathbf{V};\mathcal{G}) = b_a(\mathbf{O};P) - \chi(P;\mathcal{G}). \tag{49}$$

Suppose next that $J > 1$. Define for each $j \in \{1,\ldots,J-1\}$

$$\mathbf{I}_j \equiv \left[\mathrm{pa}_{\mathcal{G}}(W_j)\cup\{W_j\}\right]\cap \mathrm{pa}_{\mathcal{G}}(W_{j+1}).$$

Lemma 30 establishes that conditions 1) and 2) of the theorem imply that

$$\mathbf{O}\backslash\mathbf{I}_j \perp\!\!\!\perp_{\mathcal{G}} \left[\left[\mathrm{pa}_{\mathcal{G}}(W_j)\cup\{W_j\}\right]\Delta\,\mathrm{pa}_{\mathcal{G}}(W_{j+1})\right]\big|\,\mathbf{I}_j \tag{50}$$

holds for $j \in \{1,\ldots,J-1\}$. Then

$$\begin{aligned} E_P\left[b_a(\mathbf{O};P)\,|\,\mathrm{pa}_{\mathcal{G}}(W_{j+1})\right] &= E_P\left[b_a(\mathbf{O};P)\,|\,\mathrm{pa}_{\mathcal{G}}(W_{j+1})\setminus\left[\mathrm{pa}_{\mathcal{G}}(W_j)\cup\{W_j\}\right],\mathbf{I}_j\right] \\ &= E_P\left[b_a(\mathbf{O};P)\,|\mathbf{I}_j\right] \end{aligned}$$

and

$$\begin{aligned} E_P\left[b_a(\mathbf{O};P)\,|\,\mathrm{pa}_{\mathcal{G}}(W_j),W_j\right] &= E_P\left[b_a(\mathbf{O};P)\,|\,\left[\mathrm{pa}_{\mathcal{G}}(W_j)\cup\{W_j\}\right]\setminus\mathrm{pa}_{\mathcal{G}}(W_{j+1}),\mathbf{I}_j\right] \\ &= E_P\left[b_a(\mathbf{O};P)\,|\mathbf{I}_j\right]. \end{aligned}$$

Consequently

$$E_P\left[b_a(\mathbf{O};P)\,|\,\mathrm{pa}_{\mathcal{G}}(W_j),W_j\right] - E_P\left[b_a(\mathbf{O};P)\,|\,\mathrm{pa}_{\mathcal{G}}(W_{j+1})\right] = 0 \tag{51}$$

for all $j \in \{1,\ldots,J-1\}$. Thus (49) holds.

Suppose next that $K = 0$. Then condition 3) and the definition of $\mathbf{O}$ imply that

$$\{A\}\cup\mathbf{O} = \mathrm{pa}_{\mathcal{G}}(Y).$$

Consequently,

$$E_P\left[T_{P,a,\mathcal{G}} \mid Y,\mathrm{pa}_{\mathcal{G}}(Y)\right] = T_{P,a,\mathcal{G}} \text{ and } E_P\left[T_{P,a,\mathcal{G}} \mid \mathrm{pa}_{\mathcal{G}}(Y)\right] = I_a(A)b_a(\mathbf{O};P)\pi_a^{-1}(\mathbf{O};P).$$

Therefore,

$$\chi^{1,desc}_{P,a,eff}(\mathbf{V};\mathcal{G}) = \frac{I_a(A)}{\pi_a(\mathbf{O}_{min};P)}\left\{Y - b_a\left(\mathbf{O};P\right)\right\}. \tag{52}$$

Thus, since we already showed that under conditions 1) and 2), (49) holds when $J \geq 1$, we conclude that under these conditions (20) holds when $K = 0$ and $J \geq 1$.

Suppose next that $K \geq 1$. Recall that $M_{K+1} \equiv Y$. Conditions 3) and 4) of the theorem imply that

$$\{A\} \cup \mathbf{O}_{\min} \subset \mathrm{pa}_{\mathcal{G}}\left(M_k\right) \tag{53}$$

holds for $k = 1, \ldots, K+1$, and hence

$$E_P\left[T_{P,a,\mathcal{G}}|M_k, \mathrm{pa}_{\mathcal{G}}\left(M_k\right)\right] = \frac{I_a(A)}{\pi_a(\mathbf{O}_{min};P)} E_P\left[Y|M_k, \mathrm{pa}_{\mathcal{G}}\left(M_k\right)\right]. \tag{54}$$

In particular, for $k = K+1$, (54) implies

$$E_P\left[T_{P,a,\mathcal{G}}|Y, \mathrm{pa}_{\mathcal{G}}\left(Y\right)\right] = \frac{I_a(A)Y}{\pi_a(\mathbf{O}_{min};P)}. \tag{55}$$

In Lemma 31 we show that conditions 3) and 4) of the theorem imply that

$$Y \perp\!\!\!\perp_{\mathcal{G}} \left[M_{k-1}, \mathrm{pa}_{\mathcal{G}}\left(M_{k-1}\right)\right] \backslash \mathrm{pa}_{\mathcal{G}}\left(M_k\right) | \mathrm{pa}_{\mathcal{G}}\left(M_k\right), \tag{56}$$

for $k = 2, \ldots, K+1$. Then for $k = 2, \ldots, K+1$ we have

$$\begin{aligned}
&E_P\left[T_{P,a,\mathcal{G}}|M_{k-1}, \mathrm{pa}_{\mathcal{G}}\left(M_{k-1}\right)\right] = \frac{I_a(A)}{\pi_a(\mathbf{O}_{min};P)} E_P\left[Y|M_{k-1}, \mathrm{pa}_{\mathcal{G}}\left(M_{k-1}\right)\right] = \\
&\frac{I_a(A)}{\pi_a(\mathbf{O}_{min};P)} E_P\left[Y|\, \mathrm{pa}_{\mathcal{G}}\left(M_k\right), \left[M_{k-1}, \mathrm{pa}_{\mathcal{G}}\left(M_{k-1}\right)\right] \backslash \mathrm{pa}_{\mathcal{G}}\left(M_k\right)\right] = \\
&\frac{I_a(A)}{\pi_a(\mathbf{O}_{min};P)} E_P\left[Y|\, \mathrm{pa}_{\mathcal{G}}\left(M_k\right)\right] = E_P\left[T_{P,a,\mathcal{G}}|\mathrm{pa}_{\mathcal{G}}\left(M_k\right)\right],
\end{aligned}$$

where the first equality follows from (54), the second from condition 4) , the third from (56) and the fourth from (53). We therefore arrive at the conclusion that conditions 3) and 4) imply that

$$E_P\left[T_{P,a,\mathcal{G}}|M_{k-1}, \mathrm{pa}_{\mathcal{G}}\left(M_{k-1}\right)\right] - E_P\left[T_{P,a,\mathcal{G}}|\mathrm{pa}_{\mathcal{G}}\left(M_k\right)\right] = 0 \tag{57}$$

for all $k = 2, \ldots, K+1$ for all $P \in \mathcal{M}(\mathcal{G})$. We thus obtain that

$$\chi^{desc}_{P,a,eff} = E_P\left[T_{P,a,\mathcal{G}}|Y, \mathrm{pa}_{\mathcal{G}}\left(Y\right)\right] - E_P\left[T_{P,a,\mathcal{G}}|\mathrm{pa}_{\mathcal{G}}\left(M_1\right)\right]. \tag{58}$$

Next, we note that conditions 3) and 4) imply that $\mathrm{pa}_{\mathcal{G}}(M_1) = \{A\} \cup \mathbf{O}$. This implies that

$$E_P\left[T_{P,a,\mathcal{G}}|\mathrm{pa}_{\mathcal{G}}\left(M_1\right)\right] = \frac{I_a(A) b_a(\mathbf{O};P)}{\pi_a(\mathbf{O};P)}.$$

This together with (55) and (58) implies that

$$\chi^{1,desc}_{P,a,eff} = \frac{I_a(A)}{\pi_a(\mathbf{O};P)}(Y - b_a(\mathbf{O};P)). \tag{59}$$

Since we have already shown that under conditions 1) and 2) of the theorem equation (49) holds for $J \geq 1$, then adding $\chi^{non-desc}_{P,a,eff}$ and $\chi^{desc}_{P,a,eff}$ we conclude that under conditions 1)-4) of the theorem, (20) holds when $J \geq 1$ and $K \geq 1$.

Finally, when $K \geq 1$ and $J = 0$, the right hand side of (59) reduces to

$$\frac{I_a(A)}{P(A=a)}(Y - \chi_a(P;\mathcal{G})),$$

because $\mathbf{O} = \emptyset$. Thus

$$\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G}) = 0 + \chi^{1,desc}_{P,a,eff}(\mathbf{V};\mathcal{G}) = \frac{I_a(A)}{P(A=a)}(Y - \chi_a(P;\mathcal{G})) = \psi_{P,a}(\mathbf{O};\mathcal{G}).$$

This concludes the proof of the sufficiency part of the theorem.

Turn now to the proof of the necessity part of the theorem. First note that $\chi^{1,non-desc}_{P,a,eff}(\mathbf{V};\mathcal{G})$ depends on $\mathbf{V}$ only through $\mathbf{W}$, so in an abuse of notation we will write $\chi^{1,non-desc}_{P,a,eff}(\mathbf{W};\mathcal{G})$. Likewise $\chi^{1,desc}_{P,a,eff}(\mathbf{V};\mathcal{G})$ depends on $\mathbf{V}$ only through $\{A, Y\} \cup \mathbf{M} \cup \mathbf{O}$, hence we write $\chi^{1,desc}_{P,a,eff}(A, \mathbf{O}, \mathbf{M}, Y;\mathcal{G})$

Suppose first that condition 3) fails. By part 1 of Lemma 34 then there exists $P^* \in \mathcal{M}(\mathcal{G})$ such that the term $\frac{I_a(A)}{\pi_a(\mathbf{O}_{min};P^*)}Y$ does not appear in the expression for $\chi^{1,desc}_{P^*,a,eff}(A, \mathbf{O}, \mathbf{M}, Y;\mathcal{G})$. Since such term appears in the expression for $\psi_{P^*,a}(\mathbf{O};\mathcal{G})$ this shows that $\chi^1_{P^*,a,eff}(\mathbf{V};\mathcal{G}) \neq \psi_{P^*,a}(\mathbf{O};\mathcal{G})$.

Now suppose that condition 3) holds but condition 4) fails because

$$\mathrm{pa}_{\mathcal{G}}(M_{K+1}) \not\subset \mathrm{pa}_{\mathcal{G}}(M_K) \cup \{M_K\}.$$

By part 2 of Lemma 34 there exists $P^* \in \mathcal{M}(\mathcal{G})$ such that $\chi^{1,desc}_{P^*,a,eff}(A, \mathbf{O}, \mathbf{M}, Y;\mathcal{G})$ depends on M_K. Then $\chi^1_{P^*,a,eff}(\mathbf{V};\mathcal{G}) = \chi^{1,non-desc}_{P^*,a,eff}(\mathbf{W};\mathcal{G}) + \chi^{1,desc}_{P^*,a,eff}(A, \mathbf{O}, \mathbf{M}, Y;\mathcal{G})$ cannot be equal to $\psi_{P^*,a}(\mathbf{O};\mathcal{G})$, since $\psi_{P^*,a}(\mathbf{O};\mathcal{G})$ is not a function of M_K.

Next assume that condition 3) holds but condition 4) fails because

$$\mathrm{pa}_{\mathcal{G}}(M_{k^*}) \not\subset \mathrm{pa}_{\mathcal{G}}(M_{k^*-1}) \cup \{M_{k^*-1}\}$$

for some $k^* \in \{2, \ldots, K\}$ but

$$\mathrm{pa}_{\mathcal{G}}(M_k) \subset \mathrm{pa}_{\mathcal{G}}(M_{k-1}) \cup \{M_{k-1}\}$$

holds for all $k \in \{k^*+1, \ldots, K+1\}$. Then by part 3) of Lemma 34, there exists $P^* \in \mathcal{M}(\mathcal{G})$ such that

$$E_{P^*}\left[T_{P,a,\mathcal{G}} \mid M_{k^*-1}, \mathrm{pa}_{\mathcal{G}}(M_{k^*-1})\right] - E_{P^*}\left[T_{P,a,\mathcal{G}} \mid \mathrm{pa}_{\mathcal{G}}(M_{k^*})\right]$$

is a non-constant function of M_{k^*-1}. Furthermore by Lemma 31, (53) and (56) hold for all $k \in \{k^*+1, \ldots, K+1\}$. Then, arguing as above,

$$E_{P^*}\left[T_{P,a,\mathcal{G}} \mid M_{k-1}, \mathrm{pa}_{\mathcal{G}}(M_{k-1})\right] - E_{P^*}\left[T_{P,a,\mathcal{G}} \mid \mathrm{pa}_{\mathcal{G}}(M_k)\right] = 0$$

for all $k \in \{k^*+1, \ldots, K+1\}$. Also, by part 1) of Lemma 30

$$E_P\left[T_{P,a,\mathcal{G}} | Y, \mathrm{pa}_{\mathcal{G}}(Y)\right] = \frac{I_a(A)Y}{\pi_a(\mathbf{O}_{min};P)}$$

Then, with the convention that $\sum_{j=2}^{k-1}(\cdot) \equiv 0$ if $k=2$, we have

$$\chi^{1,desc}_{P^*,a,eff}(A,\mathbf{O},\mathbf{M},Y;\mathcal{G}) =$$
$$E_{P^*}\left[T_{P^*,\mathcal{G}}|M_{k^*-1},\mathrm{pa}_{\mathcal{G}}\left(M_{k^*-1}\right)\right] - E_{P^*}\left[T_{P^*,a,\mathcal{G}}|\mathrm{pa}_{\mathcal{G}}\left(M_{k^*}\right)\right] +$$
$$\frac{I_a(A)Y}{\pi_a(\mathbf{O}_{min};P^*)} + \sum_{j=2}^{k^*-1}\left\{E_{P^*}\left[T_{P^*,a,\mathcal{G}}|M_{j-1},\mathrm{pa}_{\mathcal{G}}\left(M_{j-1}\right)\right] - E_{P^*}\left[T_{P^*,a,\mathcal{G}}|\mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right]\right\} -$$
$$E_{P^*}\left[T_{P^*,a,\mathcal{G}}|\mathrm{pa}_{\mathcal{G}}\left(M_1\right)\right].$$

Now, by the topological order of $(M_1,\ldots,M_{K+1})$, M_{k^*-1} does not belong to $\mathrm{pa}_{\mathcal{G}}(M_j)$ for any $j \leq k^*-1$ and consequently none of the terms $E_{P^*}\left[T_{P^*,\mathcal{G}}|M_{j-1},\mathrm{pa}_{\mathcal{G}}\left(M_{j-1}\right)\right] - E_{P^*}\left[T_{P^*,\mathcal{G}}|\mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right]$ for $j < k^*-1$ in the last display depend on M_{k^*-1}. This then shows that $\chi^{1,desc}_{P^*,a,eff}(A,\mathbf{O},\mathbf{M},Y;\mathcal{G})$ is a non-constant function of M_{k^*-1} thus implying that $\psi_{P^*,a}\left[\mathbf{O}\left(A,Y;\mathcal{G}\right);\mathcal{G}\right] \neq \chi^{1}_{P^*,a,eff}(\mathbf{V};\mathcal{G})$ since $\psi_{P^*,a}\left[\mathbf{O}\left(A,Y;\mathcal{G}\right);\mathcal{G}\right]$ does not depend on M_{k^*-1}.

Assume now that conditions 3) and 4) hold but condition 1) fails. This can only occur if $J>1$. We have already shown that $O_T = W_J$. Then, since W_J appears only in the term $E_P\left[b_a\left(\mathbf{O};P\right)|W_J,\mathrm{pa}_{\mathcal{G}}\left(W_J\right)\right]$ of $\chi^{1,non-desc}_{P,a,eff}(\mathbf{W};\mathcal{G})$, we conclude that $\chi^{1,non-desc}_{P,a,eff}(\mathbf{W};\mathcal{G}) = g_1\left[W_J,\mathrm{pa}_{\mathcal{G}}\left(W_J\right)\right] + g_2\left(\mathbf{W}\backslash W_J\right)$ for some functions g_1 and g_2. This implies that $\chi^{1,non-desc}_{P,a,eff}(\mathbf{W};\mathcal{G})$ cannot be equal to, for instance, $b^*\left(\mathbf{O}\right) + g_2\left(\mathbf{W}\backslash O_T\right)$ for $b^*\left(\mathbf{O}\right) = O_1 \times \cdots \times O_T$. By Lemma 32 we can find $P^* \in \mathcal{M}(\mathcal{G})$ such that $b_a\left(\mathbf{O};P^*\right) = b^*\left(\mathbf{O}\right)$. For this $P^* \in \mathcal{M}(\mathcal{G})$, $\chi^{1,non-desc}_{P^*,a,eff}(\mathbf{W};\mathcal{G})$ cannot be equal to $b_a(\mathbf{O};P^*) - \chi_a(P^*;\mathcal{G})$. Then, since we have already shown that when conditions 3) and 4) of the theorem are satisfied it holds that

$$\chi^{1,desc}_{P^*,a,eff}(A,\mathbf{O},\mathbf{M},Y;\mathcal{G}) = \frac{I_a(A)}{\pi_a(\mathbf{O};P^*)}(Y - b_a(\mathbf{O};P^*)),$$

we conclude that

$$\chi^{1}_{P^*,a,eff}\left(\mathbf{V};\mathcal{G}\right) = \frac{I_a(A)}{\pi_a(\mathbf{O};P^*)}(Y - b_a(\mathbf{O};P^*)) + \chi^{1,non-desc}_{P^*,a,eff}(\mathbf{W};\mathcal{G}) \neq \psi_{P^*,a}\left(\mathbf{O};\mathcal{G}\right).$$

Assume now that conditions 1), 3) and 4) hold, but condition 2) fails because there exists $j^* \in \{2,\ldots,J-1\}$, where $J>1$, such that $\mathrm{pa}_{\mathcal{G}}(W_{j^*+1}) \setminus \{W_{j^*}\} \not\subset \mathrm{pa}_{\mathcal{G}}(W_{j^*})$. Then, by part 4) of Lemma 30 we have that

$$\mathbf{O}\backslash\mathbf{I}_{j^*} \not\perp\!\!\!\perp_{\mathcal{G}} \left[\mathrm{pa}_{\mathcal{G}}\left(W_{j^*}\right) \cup W_{j^*}\right] \triangle \mathrm{pa}_{\mathcal{G}}\left(W_{j^*+1}\right)|\mathbf{I}_{j^*}.$$

By Lemma 33, there exists $P^* \in \mathcal{M}$ such that $\chi^{1,non-desc}_{P^*,a,eff}(\mathbf{W};\mathcal{G}) = b_a(\mathbf{O};P^*) - \chi_a(P^*;\mathcal{G}) + g(\mathbf{W})$, where $g(\mathbf{W})$ is non-constant function of W_{j^*}. Then,

$$\begin{aligned}\chi^{1}_{P^*,a,eff}\left(\mathbf{V};\mathcal{G}\right) &= \chi^{1,non-desc}_{P^*,a,eff}(\mathbf{W};\mathcal{G}) + \chi^{1,desc}_{P^*,a,eff}(A,\mathbf{O},\mathbf{M},Y;\mathcal{G}) \\ &= b_a(\mathbf{O};P^*) - \chi_a(P^*;\mathcal{G}) + g(\mathbf{W}) + \frac{I_a(A)}{\pi_a(\mathbf{O}_{min};P^*)}(Y - b_a(\mathbf{O};P^*))\end{aligned}$$

cannot be equal to $\psi_{P^*,a}\left(\mathbf{O};\mathcal{G}\right) = b_a(\mathbf{O};P^*) - \chi_a(P^*;\mathcal{G}) + \frac{I_a(A)}{\pi_a(\mathbf{O}_{min};P^*)}(Y - b_a(\mathbf{O};P^*))$.

This finishes the proof of the theorem.

■

A.3 Proof of soundness of Algorithm 2 and examples

Proof that Algorithm 2 is sound for a simplified formula for $\chi^1_{P,a,eff}$

Define $\chi^{1,non-desc}_{P,a,eff}(\mathbf{V};\mathcal{G})$ and $\chi^{1,desc}_{P,a,eff}(\mathbf{V};\mathcal{G})$ as in the proof of Theorem 19.

The algorithm starts by searching for possible deletions and/or simplifications of the terms in the expression for $\chi^{1,non-desc}_{P,a,eff}$ when $J \geq 1$. If $J = 1$, we have have already shown in the proof of Theorem 19 that $\chi^{1,non-desc}_{P,a,eff}(\mathbf{V};\mathcal{G}) = b_a(\mathbf{O};P) - \chi_a(P;\mathcal{G})$. As in the proof of Theorem 19, for $J > 1$ define for each $j \in \{1,\ldots,J-1\}$

$$\mathbf{I}_j \equiv \left[\mathrm{pa}_{\mathcal{G}}(W_j) \cup \{W_j\}\right] \cap \mathrm{pa}_{\mathcal{G}}(W_{j+1}).$$

If

$$\mathbf{O}\backslash\mathbf{I}_j \perp\!\!\!\perp_{\mathcal{G}} \left[\left[\mathrm{pa}_{\mathcal{G}}(W_j) \cup \{W_j\}\right] \Delta\, \mathrm{pa}_{\mathcal{G}}(W_{j+1})\right] \big| \mathbf{I}_j \tag{60}$$

then

$$\begin{aligned} E_P\left[b_a(\mathbf{O};P) \,|\, \mathrm{pa}_{\mathcal{G}}(W_{j+1})\right] &= E_P\left[b_a(\mathbf{O};P) \,|\, \mathrm{pa}_{\mathcal{G}}(W_{j+1}) \setminus \left[\mathrm{pa}_{\mathcal{G}}(W_j) \cup \{W_j\}\right], \mathbf{I}_j\right] \\ &= E_P\left[b_a(\mathbf{O};P) \,|\mathbf{I}_j\right] \end{aligned}$$

and

$$\begin{aligned} E_P\left[b_a(\mathbf{O};P) \,|\, \mathrm{pa}_{\mathcal{G}}(W_j), W_j\right] &= E_P\left[b_a(\mathbf{O};P) \,|\, \left[\mathrm{pa}_{\mathcal{G}}(W_j) \cup \{W_j\}\right] \setminus \mathrm{pa}_{\mathcal{G}}(W_{j+1}), \mathbf{I}_j\right] \\ &= E_P\left[b_a(\mathbf{O};P) \,|\mathbf{I}_j\right]. \end{aligned}$$

Thus (60) is a graphical criterion for checking if the difference

$$E_P\left[b_a(\mathbf{O};P) \,|\, \mathrm{pa}_{\mathcal{G}}(W_j), W_j\right] - E_P\left[b_a(\mathbf{O};P) \,|\, \mathrm{pa}_{\mathcal{G}}(W_{j+1})\right] \tag{61}$$

cancel out from the expression for $\chi^1_{P,a,eff}(\mathbf{V};\mathcal{G})$ for all $P \in \mathcal{M}(\mathcal{G})$.

There is one important instance in which the graphical criterion (60) can be significantly simplified. Specifically, recall that in the proof of Theorem 19 we showed $W_J = O_T$. Suppose now that

$$\mathbf{O}\backslash O_T \subset \mathrm{pa}_{\mathcal{G}}(W_J). \tag{62}$$

Lemma 30 establishes that, under (62), the criterion (60) holds for $j = J-1$ if and only if $\mathrm{pa}_{\mathcal{G}}(W_j) \setminus \{W_{j-1}\} \subset \mathrm{pa}_{\mathcal{G}}(W_{j-1})$. Furthermore, the lemma also establishes that if for some $1 < j^* \leq J-1$

$$\mathrm{pa}_{\mathcal{G}}(W_{j+1}) \subset \mathrm{pa}_{\mathcal{G}}(W_j) \cup \{W_j\} \tag{63}$$

is valid for $j \in \{j^*,\ldots,J-1\}$, then (60) holds for $j \in \{j^*,\ldots,J-1\}$, and in addition, (60) and (63) are equivalent for $j = j^* - 1$. Note that whereas (60) requires checking d-separations, (63) requires simply checking the inclusion of sets.

Aside from the implications for term cancellations, note that when (62) holds

$$E_P\left[b_a(\mathbf{O};P) \,|W_J, \mathrm{pa}_{\mathcal{G}}(W_J)\right] = b_a(\mathbf{O};P).$$

Steps 9-26 of Algorithm 2 implement the preceding checks. Specifically, step 9 inquires if both $J > 1$ and (62) hold. If $J > 1$ but (62) does not hold, then the algorithm goes on to

inquire for each $j \in \{1, \ldots, J-1\}$, if (60) holds (see Algorithm 4) and it stores the formula

$$\chi^{1,non-desc}_{P,a,eff}(\mathbf{V};\mathcal{G}) = E_P\left[b_a(\mathbf{O};P)\,|W_J, \mathrm{pa}_{\mathcal{G}}(W_J)\right] - \chi_a(P;\mathcal{G}) + \sum_{\substack{j\in\{1,\ldots,J-1\}:\\ (60)\text{ does not hold}}} \left\{E_P\left[b_a(\mathbf{O};P)\,|\,\mathrm{pa}_{\mathcal{G}}(W_j), W_j\right] - E_P\left[b_a(\mathbf{O};P)\,|\,\mathrm{pa}_{\mathcal{G}}(W_{j+1})\right]\right\}.$$

If both $J > 1$ and (62) hold, then iteratively in reverse order from $j = J-1$, the algorithm inquires if (63) holds until the first j, if any, such that the inclusion (63) fails. If such j, say $j = j^*$ exists, j^* is necessarily greater than 1 because of the topological order of $\mathbf{W}$ and the fact that $\mathrm{irrel}(A, Y, \mathcal{G}) = \emptyset$. Then the algorithm inquires for each $j \in \{1, \ldots, j^*-1\}$ if (60) holds and it stores the formula

$$\chi^{1,non-desc}_{P,a,eff}(\mathbf{V};\mathcal{G}) = b_a(\mathbf{O};P) - \chi_a(P;\mathcal{G}) + E_P\left[b_a(\mathbf{O};P)\,|\,\mathrm{pa}_{\mathcal{G}}(W_{j^*}), W_{j^*}\right] - E_P\left[b_a(\mathbf{O};P)\,|\,\mathrm{pa}_{\mathcal{G}}(W_{j^*+1})\right] + \sum_{\substack{j\in\{1,2,\ldots,j^*-1\}:\\ (60)\text{ does not hold}}} \left\{E_P\left[b_a(\mathbf{O};P)\,|\,\mathrm{pa}_{\mathcal{G}}(W_j), W_j\right] - E_P\left[b_a(\mathbf{O};P)\,|\,\mathrm{pa}_{\mathcal{G}}(W_{j+1})\right]\right\}. \quad (64)$$

If (63) holds for all $j \in \{1, \ldots, J-1\}$ for $J > 1$ or if $J = 1$ then the algorithm stores the formula

$$\chi^{1,non-desc}_{P,a,eff}(\mathbf{V};\mathcal{G}) = b_a(\mathbf{O};P) - \chi_a(P;\mathcal{G}). \quad (65)$$

Otherwise if $J = 0$ it stores $\chi^{1,non-desc}_{P,a,eff}(\mathbf{V};\mathcal{G}) = 0$.

Importantly the expression (64) for $\chi^{1,non-desc}_{P,a,eff}(\mathbf{V};\mathcal{G})$ does not depend on the variables $\{W_{j^*+1}, \ldots, W_J\} \setminus \mathbf{O}$. Since the expression for $\chi^{1,desc}_{P,a,eff}(\mathbf{V};\mathcal{G})$ does not depend on these variables then we conclude that $\{W_{j^*+1}, \ldots, W_J\} \setminus \mathbf{O}$ do not enter into the formula for $\chi^{1}_{P,a,eff}(\mathbf{V};\mathcal{G})$ and consequently do not provide information about the parameter $\chi_a(P;\mathcal{G})$. We emphasize that this is important from a practical standpoint because if the algorithm returns expression (64), then the investigator does not need to measure these variables. A similar comment applies if the algorithm returns expression (65).

Having checked for possible simplifications of the expression of $\chi^{1,non-desc}_{P,a,eff}$, Algorithm 2 goes on to check for possible simplifications of $\chi^{1,desc}_{P,a,eff}(\mathbf{V};\mathcal{G})$. Recall that $M_{K+1} \equiv Y$.

Suppose first that $K = 0$. In the proof of Theorem 19 we showed that

$$\chi^{1,desc}_{P,a,eff}(\mathbf{V};\mathcal{G}) = \frac{I_a(A)}{\pi_a(\mathbf{O}_{min};P)}\left\{Y - b_a(\mathbf{O};P)\right\}.$$

Suppose next that $K \geq 1$. If for some $k \in \{2, \ldots, K+1\}$, it holds that

$$\{A\} \cup \mathbf{O}_{\min} \subset \mathrm{pa}_{\mathcal{G}}(M_k) \quad (66)$$

then

$$E_P\left[T_{P,a,\mathcal{G}}|M_k, \mathrm{pa}_{\mathcal{G}}(M_k)\right] = \frac{I_a(A)}{\pi_a(\mathbf{O}_{min};P)} E_P\left[Y|M_k, \mathrm{pa}_{\mathcal{G}}(M_k)\right]. \quad (67)$$

Note that for $k = K+1$, (67) is equal to

$$E_P\left[T_{P,a,\mathcal{G}}|Y, \mathrm{pa}_{\mathcal{G}}(Y)\right] = \frac{I_a(A)Y}{\pi_a(\mathbf{O}_{min}; P)}. \tag{68}$$

Now, suppose that for some $k \in \{2, \ldots, K+1\}$, in addition to (66), it holds that

$$\mathrm{pa}_{\mathcal{G}}(M_k) \subset \mathrm{pa}_{\mathcal{G}}(M_{k-1}) \cup \{M_{k-1}\} \tag{69}$$

and

$$Y \perp\!\!\!\perp_{\mathcal{G}} \left[M_{k-1}, \mathrm{pa}_{\mathcal{G}}(M_{k-1})\right] \backslash \mathrm{pa}_{\mathcal{G}}(M_k) \,|\mathrm{pa}_{\mathcal{G}}(M_k)\,. \tag{70}$$

Then, for such k, we have already shown in the proof of Theorem 19 that

$$E_P\left[T_{P,a,\mathcal{G}}|M_{k-1}, \mathrm{pa}_{\mathcal{G}}(M_{k-1})\right] - E_P\left[T_{P,a,\mathcal{G}}|\mathrm{pa}_{\mathcal{G}}(M_k)\right]$$

vanishes from the expression for $\chi^{1,desc}_{P,a,eff}(\mathbf{V}; \mathcal{G})$ for all $P \in \mathcal{M}(\mathcal{G})$.

Aside from the examination of term cancellations, we note that if

$$\mathrm{pa}_{\mathcal{G}}(M_1) = \{A\} \cup \mathbf{O} \tag{71}$$

holds, then $E_P\left[T_{P,a,\mathcal{G}}|\mathrm{pa}_{\mathcal{G}}(M_1)\right] = I_a(A)[\pi_a(\mathbf{O}_{min}; P)]^{-1} b_a(\mathbf{O}; P)$.

Steps 27-50 of Algorithm 2 implement the preceding checks. Specifically, step 27 inquires if both $K \geq 1$ and (66) hold for $k = K+1$. If $K \geq 1$ but (66) does not hold for $k = K+1$, the algorithm goes on to inquire for each $k \in \{2, \ldots, K\}$ if (66), (69) and (70) hold and subsequently if (71) holds. It then stores the formula

$$\begin{aligned}
&\chi^{1,desc}_{P,a,eff}(\mathbf{V}; \mathcal{G}) = E_P\left[T_{P,a,\mathcal{G}}|Y, \mathrm{pa}_{\mathcal{G}}(Y)\right] - E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(Y)\right] + \\
&E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(M_K), M_K\right] - \chi^{1,M_1}_{P,a,eff}(\mathbf{V}; \mathcal{G}) + \\
&\sum_{k \in \texttt{offenders_desc}(K)} \left\{E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(M_{k-1}), M_{k-1}\right] - E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(M_k)\right]\right\}
\end{aligned}$$

where

$$\chi^{1,M_1}_{P,a,eff}(\mathbf{V}; \mathcal{G}) \equiv \begin{cases} \frac{I_a(A)}{\pi_a(\mathbf{O}_{\min;P})} b_a(\mathbf{O}; P) & \text{if (71) holds} \\ E_P\left[T_{P,a,\mathcal{G}}|\mathrm{pa}_{\mathcal{G}}(M_1)\right] & \text{if (71) does not hold} \end{cases}$$

and for any $h \in \{2, \ldots, K\}$

$$\texttt{offenders\_desc}(h) \equiv \{k \in \{2, \ldots, h\} : \text{at least one of (66), (69) or (70) does not hold}\}\,.$$

See Algorithm 3. If $K \geq 1$ and (66) holds for $k = K+1$ then iteratively in reverse order from $k = K+1$ the algorithm inquires if (69) holds until the first $k \geq 2$, if any, in which the condition fails. If such k, say $k = k^*$ exists and $k^* > 2$, then it inquires for each $k \in \{2, \ldots, k^*-1\}$ if (66), (69) and (70) hold, and subsequently if (71) holds. It then stores the formula

$$\begin{aligned}
&\chi^{1,desc}_{P,a,eff}(\mathbf{V}; \mathcal{G}) = \frac{I_a(A)Y}{\pi_a(\mathbf{O}_{min}; P)} - \chi^{1,M_1}_{P,a,eff}(\mathbf{V}; \mathcal{G}) + \\
&E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(M_{k^*-1}), M_{k^*-1}\right] - E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(M_{k^*})\right] + \\
&\sum_{k \in \texttt{offenders_desc}(k^*-1)} \left\{E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(M_{k-1}), M_{k-1}\right] - E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(M_k)\right]\right\}. \qquad (72)
\end{aligned}$$

Notice that in a similar fashion as for the expression (64) for $\chi^{1,non-desc}_{P,a,eff}(\mathbf{V};\mathcal{G})$, the expression (72) does not depend on the variables $M_{k^*},\ldots,M_K$. Since the expression for $\chi^{1,non-desc}_{P,a,eff}$ does not depend on these variables, we conclude that $M_{k^*},\ldots,M_K$ do not enter into the formula for $\chi^{1}_{P,a,eff}(\mathbf{V};\mathcal{G})$ and consequently do not provide information about the parameter $\chi_a(P;\mathcal{G})$.

If $k^*=2$, then it stores

$$\chi^{1,desc}_{P,a,eff}(\mathbf{V};\mathcal{G}) = \frac{I_a(A)Y}{\pi_a(\mathbf{O}_{min};P)} - \chi^{1,M_1}_{P,a,eff}(\mathbf{V};\mathcal{G}) + E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}\left(M_{k^*-1}\right),M_{k^*-1}\right] - E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}\left(M_{k^*}\right)\right]. \tag{73}$$

If no such k^* exists condition (71) automatically holds. Then the algorithm stores the formula

$$\chi^{1,desc}_{P,a,eff}(\mathbf{V};\mathcal{G}) = \frac{I_a(A)}{\pi_a(\mathbf{O}_{min};P)}\left\{Y - b_a\left(\mathbf{O};P\right)\right\}. \tag{74}$$

If $K=0$ then the algorithm also stores the formula in (74). This concludes the proof.

Example 11 *Consider the DAG in Figure 10 c). In this DAG,* $\mathbf{O}=\mathbf{O}_{min}=\{O_T\}=\{W_5\}$ *with* $T=1$. *Therefore condition* (62) *holds trivially. However, condition* (63) *with* $j=4$ *fails, because* W_2 *is a parent of* W_5 *but not of* W_4. *The algorithm now goes on to check condition* (60) *for each* $j=1,2,3,4$. *The following table lists the results.*

j	$\mathbf{I}_j$	$\left[\mathrm{pa}_{\mathcal{G}}(W_j)\cup\{W_j\}\triangle\mathrm{pa}_{\mathcal{G}}(W_{j+1})\right]$	$\mathbf{O}\setminus\mathbf{I}_j$	(60)
1	W_1	$\emptyset$	W_5	holds
2	W_2	W_1	W_5	holds
3	W_3	W_2	W_5	fails
4	W_4	$\{W_2,W_3\}$	W_5	fails

The algorithm then stores the formula

$$\begin{aligned}\chi^{1,non-desc}_{P,a,eff}(\mathbf{V};\mathcal{G}) &= b_a(W_5;P) - \chi_a(P;\mathcal{G}) \\ &\quad + E_P\left[b_a(W_5;P)\mid W_4,\mathrm{pa}_{\mathcal{G}}(W_4)\right] - E_P\left[b_a(W_5;P)\mid \mathrm{pa}_{\mathcal{G}}(W_5)\right] \\ &\quad + E_P\left[b_a(W_5;P)\mid W_3,\mathrm{pa}_{\mathcal{G}}(W_3)\right] - E_P\left[b_a(W_5;P)\mid W_3\right] \\ &= b_a(W_5;P) - \chi_a(P;\mathcal{G}) + E_P\left[b_a(O;P)\mid W_3,W_4\right] - E_P\left[b_a(W_5;P)\mid W_2,W_4\right] \\ &\quad + E_P\left[b_a(W_5;P)\mid W_2,W_3\right] - E_P\left[b_a(W_5;P)\mid W_3\right].\end{aligned}$$

This example illustrates the following interesting points.

For $j=2$ *the d-separation* (60) *holds and consequently the term* (61) *vanishes from the expression for* $\chi^{1,non-desc}_{P,a,eff}$. *However,* W_2 *appears in the expression for* $\chi^{1,non-desc}_{P,a,eff}$ *and therefore it appears also in the expression for the efficient influence function* $\chi^{1}_{P,a,eff}$. *Thus,* W_2 *provides information about* $\chi_a(P;\mathcal{G})$ *even though the term* (61) *vanishes for* $j=2$. *In contrast, for* $j=1$ *term* (61) *vanishes and* W_1 *does not enter into the expression for* $\chi^{1,non-desc}_{P,a,eff}$. *This illustrates the point that once condition* (63) *fails, the check of the d-separation condition* (60) *is useful for detecting term cancellations but not for deciding if the corresponding node is informative about the parameter* $\chi_a(P;\mathcal{G})$.

Another interesting point illustrated by this example is that the composition of the set $\text{pa}_{\mathcal{G}}(A)$ *does not affect the expression for* $\chi^{1,non-desc}_{P,a,eff}$*. That is, all or a subset of the orange edges could have been absent in the DAG and nevertheless the expression for* $\chi^{1,non-desc}_{P,a,eff}$ *would have remained the same. However, which elements of* $\mathbf{W}$ *are members of the set* $\text{pa}_{\mathcal{G}}(A)$ *does affect the composition of the minimal optimal adjustment set* $\mathbf{O}_{min}$*. For instance in the DAG of Figure 10 c),* $\mathbf{O}_{min} = \mathbf{O} = \{W_5\}$*. Instead, if all the orange arrows had been absent, then* $\mathbf{O}_{min}$ *would have been empty.*

In this DAG, $\mathbf{M} = \emptyset$ *and hence* $K = 0$*. The algorithm then stores the formula in* (74) *and finally returns*

$$\chi^{1}_{P,a,eff}(\mathbf{V};\mathcal{G}) = b_a(W_5;P) - \chi_a(P;\mathcal{G}) + E_P\left[b_a(W_5;P) \mid W_3, W_4\right] - E_P\left[b_a(W_5;P) \mid W_2, W_4\right] + E_P\left[b_a(W_5;P) \mid W_2, W_3\right] - E_P\left[b_a(W_5;P) \mid W_3\right] + \frac{I_a(A)}{\pi_a(W_5;P)}\left\{Y - b_a\left(W_5;P\right)\right\}.$$

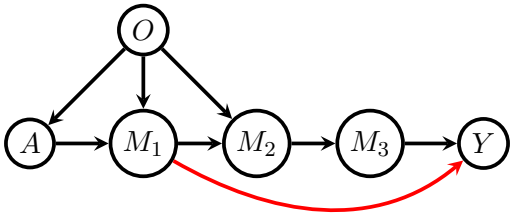

Figure 13: A DAG where the NP-**O** estimator is inefficient.

Example 12 *Consider the DAG in Figure 13. In this case* $\mathbf{O} = \mathbf{O}_{min} = \{O\} \equiv \{W_1\}$*,* $J = T = 1$*,* $\mathbf{M} = \{M_1, M_2, M_3\}$ *and* $K = 3$*. Because* $J = 1$ *the algorithm stores the formula* (65)*. In addition, it is easy to check that, conditions* (66)*,* (69) *and* (70) *hold for* $k = 2$*, hence*

$$E_P\left[T_{P,a,\mathcal{G}} \mid M_1, \text{pa}_{\mathcal{G}}(M_1)\right] - E_P\left[T_{P,a,\mathcal{G}} \mid \text{pa}_{\mathcal{G}}(M_2)\right] \tag{75}$$

vanishes for all $P \in \mathcal{M}(\mathcal{G})$*. On the other hand, condition* (66) *fails for* $k = 4$*, condition* (70) *fails for* $k = 3$ *and* $\{A, \mathbf{O}\} = \text{pa}_{\mathcal{G}}(M_1)$*. Hence the algorithm stores*

$$\begin{aligned}\chi^{1,non-desc}_{P,a,eff}(\mathbf{V};\mathcal{G}) = {} & E_P\left[T_{P,a,\mathcal{G}} \mid Y, \text{pa}_{\mathcal{G}}(Y)\right] - E_P\left[T_{P,a,\mathcal{G}} \mid \text{pa}_{\mathcal{G}}(Y)\right] \\ & + E_P\left[T_{P,a,\mathcal{G}} \mid M_3, \text{pa}_{\mathcal{G}}(M_3)\right] - E_P\left[T_{P,a,\mathcal{G}} \mid \text{pa}_{\mathcal{G}}(M_3)\right] \\ & + E_P\left[T_{P,a,\mathcal{G}} \mid M_2, \text{pa}_{\mathcal{G}}(M_2)\right] - \frac{I_a(A)b_a(O;P)}{\pi_a(O;P)}.\end{aligned}$$

Notice that even though (75) *vanishes,* $\chi^{1,non-desc}_{P,a,eff}(\mathbf{V};\mathcal{G})$ *depends on* M_1 *because*

$$E_P\left[T_{P,a,\mathcal{G}} \mid Y, \text{pa}_{\mathcal{G}}(Y)\right] = Y E_P\left[\frac{I_a(A)b_a(O;P)}{\pi_a(O;P)} \mid M_1, M_3\right].$$

A.4 Auxiliary results

Lemma 27 *If* $\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{Z}_1 \backslash \mathbf{Z}_2 \mid \mathbf{Z}_2$ *then* $E_P\left[\frac{1}{\pi_{\mathbf{a}}(\mathbf{Z}_2;P)}\middle|\mathbf{A}=\mathbf{a},\mathbf{Z}_1\right] = \frac{1}{\pi_{\mathbf{a}}(\mathbf{Z}_1;P)}$ *for all* $P \in \mathcal{M}(\mathcal{G})$.

Proof [Proof of Lemma 27]

$$\begin{aligned} & E_P\left[\frac{1}{\pi_{\mathbf{a}}(\mathbf{Z}_2;P)}\middle|\mathbf{A}=\mathbf{a},\mathbf{Z}_1\right]\pi_{\mathbf{a}}(\mathbf{Z}_1;P) \equiv E_P\left[\frac{1}{\pi_{\mathbf{a}}(\mathbf{Z}_2;P)}\middle|\mathbf{A}=\mathbf{a},\mathbf{Z}_1\right]P(\mathbf{A}=\mathbf{a}|\mathbf{Z}_1) \\ = \; & E_P\left[\frac{I_{\mathbf{a}}(\mathbf{A})}{\pi_{\mathbf{a}}(\mathbf{Z}_2;P)}\middle|\mathbf{Z}_1\right] = E_P\left[\frac{E_P(I_{\mathbf{a}}(\mathbf{A})|\mathbf{Z}_2,\mathbf{Z}_1)}{\pi_{\mathbf{a}}(\mathbf{Z}_2;P)}\middle|\mathbf{Z}_1\right] = 1 \end{aligned}$$

where the last equality follows from the fact that, $\mathbf{A} \perp\!\!\!\perp \mathbf{Z}_1 \backslash \mathbf{Z}_2 \mid \mathbf{Z}_2\ [P]$ implies $E_P(I_{\mathbf{a}}(\mathbf{A})|\mathbf{Z}_2,\mathbf{Z}_1) = E_P(I_{\mathbf{a}}(\mathbf{A})|\mathbf{Z}_2) = \pi_{\mathbf{a}}(\mathbf{Z}_2;P)$. ■

Lemma 28 *If* $\mathbf{Z}$ *is a minimal adjustment set relative to* (A,Y) *in DAG* $\mathcal{G}$*, then for all* W *in* $\mathbf{Z}$ *there exists a path* δ *between* W *and* A *that is open given* $\mathbf{Z}\backslash W$.

Proof [Proof of Lemma 28] Since $\mathbf{Z}$ is a minimal adjustment set, we know (see Shpitser et al. (2010)) that there exists a non-causal γ path between A and Y that is open when we condition on $\mathbf{Z}\backslash W$ but is blocked when we condition on $\mathbf{Z}$. The path γ must intersect W because if it did not, since the path is open when we condition on $\mathbf{Z}\backslash W$ it would also be open when we condition on $\mathbf{Z}$. Let δ be the subpath of γ that goes from A to the first ocurrence of W in γ. δ is open given $\mathbf{Z}\backslash W$, since γ is open given $\mathbf{Z}\backslash W$. ■

Lemma 29 *Let* $V \in \mathrm{dir}(A,Y,\mathcal{G})$ *and* $W \in \mathrm{indir}(A,Y,\mathcal{G})$. *Then* $V \perp\!\!\!\perp_{\mathcal{G}} W \mid A, \mathbf{F}(A,Y,\mathcal{G})$.

Proof [Proof of Lemma 29] Let $\mathbf{F} \equiv \mathbf{F}(A,Y,\mathcal{G})$. We will show that no path between V and W can be open given $A,\mathbf{F}$. We analyze separately paths that (i) are directed, (ii) are not directed and have exactly one fork and (iii) are not directed and have at least one collider. We use the notation $T \rightrightarrows S$ to represent a directed path between T and S.

(i) Directed

Assume that there is a directed path between V and W and call it δ. Assume first that δ leaves V through the front-door. If $V = Y$, since W is an ancestor of A, this implies that Y is an ancestor of A, which is a contradiction. If $V \neq Y$, since V has a directed path to Y that does not intersect A, we deduce that $V \in \mathbf{F}$, a contradiction. Assume now that δ leaves V through the backdoor. This implies that there is a directed path betweeen W and Y that does not intersect A, which is a contradiction.

Hence, there are no directed paths between V and W that are open given $(A,\mathbf{F})$.

(ii) Not directed, exactly one fork

Assume there is a path between V and W that has at exactly one fork, and consequently no colliders, and is open given $(A,\mathbf{F})$. Call the path δ and call the fork, H. Recall that W is an ancestor of A. Since V is either equal to Y or has a directed path to Y that does not

intersect A, the path $V \leftleftarrows H \rightrightarrows W \rightrightarrows A$ shows that $H \in \mathbf{F}$ and hence δ is blocked by $\mathbf{F}$, a contradiction.

(iii) Not directed, with at least one collider

Assume there is a path between V and W that has at least one collider and is open given $(A, \mathbf{F})$. Call the path δ. All colliders in δ must be either in $(A, \mathbf{F})$ or have a descendant in $(A, \mathbf{F})$. Hence, all colliders are ancestors of A.

Assume first that δ leaves V through the frontdoor. Consider the collider in δ that is closest to V and call it C. If $V = Y$, then the directed path $Y \rightrightarrows C \rightrightarrows A$ shows that A is a descendant of Y, a contradiction. If $V \neq Y$, since V has a directed path to Y that does not intersect A, the path $Y \leftleftarrows V \rightrightarrows C \rightrightarrows A$ shows that $V \in \mathbf{F}$, which is a contradiction.

Assume now that δ leaves V throught the backdoor. Consider the collider in δ that is closest to V and call it D. Because in the subpath of δ between V and D the edge with endpoint V points into V and the edge with endpoint D points to D then in that subpath there has to be a fork, say K. Such K belongs to $\mathbf{F}$, because K has directed path to D and D is an ancestor of A and also K has a directed path to V that does not intersect A and V is either equal to Y or has directed path to Y that does not intersect A. Hence δ is blocked by K, which is a contradiction. This concludes the proof of the lemma. ■

Lemma 30 *Assume that $\mathcal{G}$ is a DAG and A and Y are two distinct vertices in $\mathcal{G}$ such that $A \in \mathrm{an}_{\mathcal{G}}(Y)$. Let $\mathbf{W} \equiv \mathrm{de}^{c}_{\mathcal{G}}(A)$ and $\mathbf{O} \equiv \mathbf{O}(A, Y, \mathcal{G})$. Assume that $\mathrm{irrel}(A, Y, \mathcal{G}) = \emptyset$. Write $\mathbf{W} \equiv (W_1, \ldots, W_J)$, where we assume $J \geq 1$ and write $\mathbf{O} \equiv (O_1, \ldots, O_T)$ in topological order relative to $\mathcal{G}$. Assume $\mathbf{O} \backslash O_T \subset \mathrm{pa}_{\mathcal{G}}(O_T)$. For $j \in \{1, \ldots, J-1\}$ let $\mathbf{I}_j \equiv \left[\mathrm{pa}_{\mathcal{G}}(W_j) \cup \{W_j\}\right] \cap \mathrm{pa}_{\mathcal{G}}(W_{j+1})$. Then,*

1. *$W_J = O_T$ and, if $J \geq 2, W_{J-1} \in \mathrm{pa}_{\mathcal{G}}(W_J)$.*

2. *Suppose $J \geq 2$. If for some $1 < j^* \leq J-1$ it holds that for $j \in \{j^*, \ldots, J-1\}$,*

$$\mathrm{pa}_{\mathcal{G}}(W_{j+1}) \setminus \{W_j\} \subset \mathrm{pa}_{\mathcal{G}}(W_j), \tag{76}$$

then

$$W_j \in \mathrm{pa}_{\mathcal{G}}(W_{j+1}) \text{ for } j \in \{j^* - 1, j^*, \ldots, J-1\} \tag{77}$$

and

$$\mathbf{O} \backslash \mathbf{I}_j \perp\!\!\!\perp_{\mathcal{G}} \left[\mathrm{pa}_{\mathcal{G}}(W_j) \cup W_j\right] \triangle \mathrm{pa}_{\mathcal{G}}(W_{j+1}) \,|\mathbf{I}_j \text{ for } j \in \{j^*, \ldots, J-1\} \tag{78}$$

3. *Suppose $J \geq 2$ and that for some $j^* \in \{2, \ldots, J-1\}$ it holds that*

$$\mathrm{pa}_{\mathcal{G}}(W_{j^*+1}) \setminus \{W_{j^*}\} \not\subset \mathrm{pa}_{\mathcal{G}}(W_{j^*}) \tag{79}$$

and that (76) holds for all $j \in \{j^* + 1, \ldots, J-1\}$ if $j^* < J-1$. Then,

$$\mathbf{O} \backslash \mathbf{I}_{j^*} \not\perp\!\!\!\perp_{\mathcal{G}} \left[\mathrm{pa}_{\mathcal{G}}(W_{j^*}) \cup W_{j^*}\right] \triangle \mathrm{pa}_{\mathcal{G}}(W_{j^*+1}) \,|\mathbf{I}_{j^*}. \tag{80}$$

Proof To prove 1), note that, since $\text{irrel}(A, Y, \mathcal{G}) = \emptyset$, there exists a directed path between W_J and Y that does not intersect A. Let W be a child of W_J in that path. Then W cannot be in the set $\{W_1, \ldots, W_J\}$ because W_J is the last element in the topolocally ordered sequence $W_1, \ldots, W_J$ of non-descendants of A. Then $W \in \mathbf{M} \cup \{Y\}$ which implies that $W_J \in \mathbf{O}$ and, since $(O_1, \ldots, O_T)$ is ordered topologically, we conclude that $W_J = O_T$. Next, suppose $J \geq 2$ and that $W_{J-1} \notin \text{pa}_{\mathcal{G}}(W_J)$. Then, $W_{J-1} \notin \mathbf{O}$ because by assumption, $\mathbf{O} \backslash O_T \subset \text{pa}_{\mathcal{G}}(W_J)$. This implies that W_{J-1} is either an ancestor of Y such that all the directed paths between W_{J-1} and Y intersect A, or W_{J-1} is not an ancestor of Y. Both possibilities are impossible because they contradict that $\text{irrel}(A, Y, \mathcal{G}) = \emptyset$.

Turn now to the proof of part 2). We will first show (77) by reverse induction on j^*. Suppose $j^* = J - 1$. We want to show that $W_{J-2} \in \text{pa}_{\mathcal{G}}(W_{J-1})$. If $W_{J-2} \in \mathbf{O}$ then by $\mathbf{O} \backslash O_T \subset \text{pa}_{\mathcal{G}}(O_T)$ and part 1) of this lemma, $W_{J-2} \in \text{pa}_{\mathcal{G}}(W_J)$, which then implies by (76) applied to $j = J - 1$ that $W_{J-2} \in \text{pa}_{\mathcal{G}}(W_{J-1})$. Suppose next that $W_{J-2} \notin \mathbf{O}$ and $W_{J-2} \notin \text{pa}_{\mathcal{G}}(W_{J-1})$, then by (76), $W_{J-2} \notin \text{pa}_{\mathcal{G}}(W_J)$. Consequently, W_{J-2} is either an ancestor of Y such that all the directed paths between W_{J-2} and Y intersect A or W_{J-2} is not an ancestor of Y. Both possibilities are impossible because they contradict that $\text{irrel}(A, Y, \mathcal{G}) = \emptyset$. This shows that (77) is true for $j^* = J - 1$. Suppose now that the result holds for $j^* = m, \ldots, J - 1$, for some $2 < m \leq J - 1$. We will show that it also holds for $j^* = m - 1$. Henceforth suppose that (76) holds for $j \in \{m - 1, \ldots, J - 1\}$. Then, (76) holds for $j \in \{m, \ldots, J - 1\}$ and consequently, by the inductive hypothesis, (77) holds for $j \in \{m - 1, m, \ldots, J - 1\}$. It remains to show that $W_{m-2} \in \text{pa}_{\mathcal{G}}(W_{m-1})$. Suppose that $W_{m-2} \in \mathbf{O}$, then by $\mathbf{O} \backslash O_T \subset \text{pa}_{\mathcal{G}}(O_T)$ and part 1), $W_{m-2} \in \text{pa}_{\mathcal{G}}(W_J)$, which then implies, by (76) being valid for all $j \in \{m - 1, \ldots, J - 1\}$, that $W_{m-2} \in \text{pa}_{\mathcal{G}}(W_J) \setminus \{W_{m-1}, \ldots, W_{J-1}\} \subset \text{pa}_{\mathcal{G}}(W_{J-1}) \setminus \{W_{m-1}, \ldots, W_{J-2}\} \subset \cdots \subset \text{pa}_{\mathcal{G}}(W_{m-1})$. On the other hand, if $W_{m-2} \notin \mathbf{O}$, since $\text{irrel}(A, Y, \mathcal{G}) = \emptyset$, necessarily $W_{m-2} \in \text{pa}_{\mathcal{G}}(W_j)$ for some $j > m - 2$. Arguing as before, this implies that $W_{m-2} \in \text{pa}_{\mathcal{G}}(W_{m-1})$.

Next we prove (78). Suppose that for $j \in \{j^*, \ldots, J - 1\}$, (76) holds. Then, for $j \in \{j^*, \ldots, J - 1\}$ we have $\mathbf{I}_j = \left[\text{pa}_{\mathcal{G}}(W_j) \cup \{W_j\}\right] \cap \left[\left[\text{pa}_{\mathcal{G}}(W_{j+1}) \setminus \{W_j\}\right] \cup \{W_j\}\right] = \left[\left[\text{pa}_{\mathcal{G}}(W_{j+1}) \setminus \{W_j\}\right] \cup W_j\right] = \text{pa}_{\mathcal{G}}(W_{j+1})$, where the first and third equalities follow by (77) and the second follows by (76). On the other hand, because by assumption $\mathbf{O} \backslash O_T \subset \text{pa}_{\mathcal{G}}(W_J)$, then $\mathbf{O} \setminus (W_{j+1}, \ldots, W_J) \subset \text{pa}_{\mathcal{G}}(W_{j+1})$. Consequently, (78) holds if and only if

$$\mathbf{O} \cap (W_{j+1}, \ldots, W_J) \perp\!\!\!\perp_{\mathcal{G}} \text{pa}_{\mathcal{G}}(W_j) \setminus \text{pa}_{\mathcal{G}}(W_{j+1}) \,|\, \text{pa}_{\mathcal{G}}(W_{j+1}). \tag{81}$$

We will show by contradiction that (81) holds, and consequently that (78) holds, for $j \in \{j^*, j^* + 1, \ldots, J - 1\}$. Suppose that (81) were not true for some $j \in \{j^*, j^* + 1, \ldots, J - 1\}$. Then there would exist $u \geq j + 1$ and $l < j$ such that $W_u \not\perp\!\!\!\perp_{\mathcal{G}} W_l | \text{pa}_{\mathcal{G}}(W_{j+1})$ with $W_u \in \mathbf{O}$ and $W_l \in \text{pa}_{\mathcal{G}}(W_j) \setminus \text{pa}_{\mathcal{G}}(W_{j+1})$. Because, by (77), $W_j \in \text{pa}_{\mathcal{G}}(W_{j+1})$, the path between W_l and W_u that would be open given $\text{pa}_{\mathcal{G}}(W_{j+1})$ would necessarily have to include an edge $W_{l^*} \to W_{u^*}$ for some $l^* < j$ and $u^* \geq j + 1$. If $u^* = j + 1$, then this implies that $W_{l^*} \in \text{pa}_{\mathcal{G}}(W_{j+1})$ which is impossible because it contradicts $W_u \not\perp\!\!\!\perp_{\mathcal{G}} W_l | \text{pa}_{\mathcal{G}}(W_{j+1})$. If $u^* > j+1$, then by $W_{l^*} \in \text{pa}_{\mathcal{G}}(W_{u^*})$ we have $W_{l^*} \in \text{pa}_{\mathcal{G}}(W_{u^*}) \setminus \{W_{j+1}, \ldots, W_{u^*-1}\}$ because $l^* < j$. However, by (76), $\text{pa}_{\mathcal{G}}(W_{u^*}) \setminus \{W_{j+1}, \ldots, W_{u^*-1}\} \subset \text{pa}_{\mathcal{G}}(W_{u^*-1}) \setminus \{W_{j+1}, \ldots, W_{u^*-2}\} \subset \cdots \subset \text{pa}_{\mathcal{G}}(W_{j+1})$ which then implies that $W_{l^*} \in \text{pa}_{\mathcal{G}}(W_{j+1})$ again contradicting $W_u \not\perp\!\!\!\perp_{\mathcal{G}} W_l | \text{pa}_{\mathcal{G}}(W_{j+1})$. This proves (78).

Turn now to the proof of part 4). Suppose that $\mathrm{pa}_{\mathcal{G}}(W_{j^*+1}) \setminus \{W_{j^*}\} \not\subset \mathrm{pa}_{\mathcal{G}}(W_{j^*})$ and that (76) holds for all $j \in \{j^*+1, \ldots, J-1\}$ if $j^* < J-1$. Then there exists $l < j^*$ such that $W_l \in \mathrm{pa}_{\mathcal{G}}(W_{j^*+1}) \setminus \mathrm{pa}_{\mathcal{G}}(W_{j^*})$. By (77), $W_j \in \mathrm{pa}_{\mathcal{G}}(W_{j+1})$ for all $j = j^*, j^*+1, \ldots, J-1$. Consequently, the path $W_l \to W_{j^*+1} \to W_{j^*+2} \to \circ\cdots\circ \to W_J$ is open in $\mathcal{G}$ when conditioning on $\mathbf{I}_{j^*}$. By part 1), $W_J = O_T \in \mathbf{O} \cap (W_{j+1}, \ldots, W_J)$, and $W_l \in \left[\mathrm{pa}_{\mathcal{G}}(W_{j^*}) \cup W_{j^*}\right] \triangle \mathrm{pa}_{\mathcal{G}}(W_{j^*+1})$, thus the aforementioned open path shows that (80) holds. This concludes the proof of (79). ■

Lemma 31 *Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$ and let A and Y be two distinct vertices in $\mathbf{V}$. Suppose that* $\mathrm{irrel}(A, Y; \mathcal{G}) = \emptyset$. *Suppose* $\mathbf{M} \equiv \mathrm{de}_{\mathcal{G}}(A) \setminus \{A, Y\} \neq \emptyset$ *and let* $(M_1, \ldots, M_K)$ *be the elements of* $\mathbf{M}$ *sorted topologically. Let* $M_0 \equiv A$ *and* $M_{K+1} \equiv Y$.

Suppose that for some $k^* \geq 2$, the following inclussion holds for $k \in \{k^*, \ldots, K+1\}$

$$\mathrm{pa}_{\mathcal{G}}(M_k) \subset \mathrm{pa}_{\mathcal{G}}(M_{k-1}) \cup \{M_{k-1}\}. \tag{82}$$

Then, $M_K \in \mathrm{pa}_{\mathcal{G}}(Y)$ and for all $k \in \{k^*, \ldots, K+1\}$

(i) $M_{k-2} \in \mathrm{pa}(M_{k-1})$ and

(ii) $Y \perp\!\!\!\perp_{\mathcal{G}} \left[M_{k-1}, \mathrm{pa}_{\mathcal{G}}(M_{k-1})\right] \setminus \mathrm{pa}_{\mathcal{G}}(M_k) \,|\, \mathrm{pa}_{\mathcal{G}}(M_k)$.

Proof

That $M_K \in \mathrm{pa}_{\mathcal{G}}(Y)$ follows from $\mathrm{irrel}(A, Y; \mathcal{G}) = \emptyset$ and the fact that M_K is last in the topological order of $\mathbf{M}$.

To show (i), assume that for some $k^* \geq 2$, (82) holds for all $k \in \{k^*, \ldots, K+1\}$. Let $k \in \{k^*, \ldots, K+1\}$. If $k = k^* = 2$, then $M_{k-2} = A \in \mathrm{pa}_{\mathcal{G}}(M_1)$ for otherwise M_1 would not be a descendant of A. Next assume $k > 2$. The assumption that $\mathrm{irrel}(A, Y; \mathcal{G}) = \emptyset$ and the topological order of $(M_1, \ldots, M_K)$ implies that $M_{k-2} \in \mathrm{pa}_{\mathcal{G}}(M_r)$ for some $r \in \{k-1, k, \ldots, K+1\}$. If $r = k-1$ we are done. If $r \geq k$, then $r \in \{k^*, \ldots, K+1\}$ and consequently (82) implies that $\mathrm{pa}_{\mathcal{G}}(M_r) \subset \mathrm{pa}_{\mathcal{G}}(M_{r-1}) \cup \{M_{r-1}\} \subset \cdots \subset \mathrm{pa}_{\mathcal{G}}(M_{k-1}) \cup \{M_{r-1}, M_{r-2}, \ldots, M_{k-1}\}$ Consequently, $M_{k-2} \in \mathrm{pa}_{\mathcal{G}}(M_{k-1}) \cup \{M_{r-1}, M_{r-2}, \ldots, M_{k-1}\}$ and since $M_{k-2} \notin \{M_{r-1}, M_{r-2}, \ldots, M_{k-1}\}$ then $M_{k-2} \in \mathrm{pa}_{\mathcal{G}}(M_{k-1})$.

To show (ii), assume that for some $k^* \geq 2$, (82) holds for all $k \in \{k^*, \ldots, K+1\}$. Let $k \in \{k^*, \ldots, K+1\}$. Assumption (82) implies that

$$\begin{aligned} \mathrm{pa}_{\mathcal{G}}(Y) &\subset \mathrm{pa}_{\mathcal{G}}(M_K) \cup \{M_K\} \subset \mathrm{pa}_{\mathcal{G}}(M_{K-1}) \cup \{M_K, M_{K-1}\} \subset \cdots \\ &\subset \mathrm{pa}_{\mathcal{G}}(M_k) \cup \{M_K, M_{K-1}, \ldots, M_k\} \subset \mathrm{pa}_{\mathcal{G}}(M_{k-1}) \cup \{M_K, M_{K-1}, \ldots, M_{k-1}\} \end{aligned} \tag{83}$$

By part (i) we have $M_{k-1} \in \mathrm{pa}_{\mathcal{G}}(M_k)$. Then, assertion (ii) is the same as $Y \perp\!\!\!\perp_{\mathcal{G}} \mathrm{pa}_{\mathcal{G}}(M_{k-1}) \setminus \mathrm{pa}_{\mathcal{G}}(M_k) \,|\, \mathrm{pa}_{\mathcal{G}}(M_k)$ Suppose the latter d-separation is false. Let $M_j \in \mathrm{pa}_{\mathcal{G}}(M_{k-1}) \setminus \mathrm{pa}_{\mathcal{G}}(M_k)$ such that $Y \not\perp\!\!\!\perp_{\mathcal{G}} M_j | \mathrm{pa}_{\mathcal{G}}(M_k)$. Because Y has no descendants in the DAG, then any open path between M_j and Y must end with an edge pointing into Y. If such path is open when we condition on $\mathrm{pa}_{\mathcal{G}}(M_k) = \left[\mathrm{pa}_{\mathcal{G}}(Y) \setminus \{M_K, M_{K-1}, \ldots, M_k\}\right] \cup \left[\mathrm{pa}_{\mathcal{G}}(M_k) \setminus \mathrm{pa}_{\mathcal{G}}(Y)\right]$, then this edge must connect a vertex $M_t \in \{M_K, M_{K-1}, \ldots, M_k\}$ with Y. This is because any other vertex would be in $\mathrm{pa}_{\mathcal{G}}(Y) \setminus \{M_K, M_{K-1}, \ldots, M_k\}$ and the path would then be closed because we are conditioning on $\mathrm{pa}_{\mathcal{G}}(Y) \setminus \{M_K, M_{K-1}, \ldots, M_k\}$ .Then the path between M_j and Y that is open when we condition on $\mathrm{pa}_{\mathcal{G}}(M_k)$ must be of the form

$$M_j - \circ - \circ \cdots \circ - V \to M_t \to Y \tag{84}$$

or

$$M_j - \circ - \circ \cdots \circ - V \leftarrow M_t \rightarrow Y \tag{85}$$

for some $t \in \{k, k+1, \ldots, K\}$ and some $V \in \mathbf{V}$. We now argue that it cannot be of the form (85). Suppose it was of the form (85). Then, V would belong to $\mathbf{M}$ because V is a child of a descendant of A and consequently it is itself a descendant of A. By the topological order of $(M_1, \ldots, M_K)$ this would imply that $V = M_h$ for some $h > t$. But in such case the path between M_j and M_h would eventually intersect a collider M_r for some $r > h$, i.e. it would be of the form $M_j - \circ - \circ \cdots \circ - \rightarrow M_r \leftarrow \circ \cdots \leftarrow \circ \leftarrow M_h \leftarrow M_t \rightarrow Y$. However, this is impossible because by $r > h > t \in \{k, k+1, \ldots, K\}$ we have that neither M_r nor its descendants are in $\mathrm{pa}_{\mathcal{G}}(M_k)$, so the path is closed at the collider M_r when we condition on $\mathrm{pa}_{\mathcal{G}}(M_k)$.

We thus conclude that if an open path exists it must be of the form (84) for some $t \in \{k, k+1, \ldots, K\}$. However, we will now show that this is also impossible. First we note that the assumption that the path is open when we condition on $\mathrm{pa}_{\mathcal{G}}(M_k)$ implies that $V \notin \mathrm{pa}_{\mathcal{G}}(M_k)$.This implies that $k+1 \leq t \leq K$. Next, note that because $V \in \mathrm{pa}_{\mathcal{G}}(M_t)$ and $\mathrm{pa}_{\mathcal{G}}(M_t) \subset \mathrm{pa}_{\mathcal{G}}(M_k) \cup \{M_k, \ldots, M_{t-1}\}$ this implies that $V \in \{M_k, \ldots, M_{t-1}\}$. So, we conclude that the open path must be of the form

$$M_j - \circ - \circ \cdots \circ - V' \rightarrow M_h \rightarrow M_t \rightarrow Y \tag{86}$$

or

$$M_j - \circ - \circ \cdots \circ - V' \leftarrow M_h \rightarrow M_t \rightarrow Y \tag{87}$$

for some $k \leq h < t \leq K$. However, reasoning as above we rule out the path (87) and conclude that the path must be of the form (86) for $V' = M_r$ with r such that $k \leq r < h < t \leq K$. Continuing in this fashion we arrive at the conclusion that the path must be of the form $M_j - \circ - \circ \cdots \circ - V^* \rightarrow M_k \ldots M_r \rightarrow M_h \rightarrow M_t \rightarrow Y$. But this contradicts the assumption that the path is open when we condition on $\mathrm{pa}_{\mathcal{G}}(M_k)$ since $V^* \in \mathrm{pa}_{\mathcal{G}}(M_k)$. This concludes the proof. ■

Lemma 32 *Assume that $\mathcal{G}$ is a DAG and A and Y are two distinct vertices in $\mathcal{G}$ such that $A \in \mathrm{an}_{\mathcal{G}}(Y)$ and $\mathrm{irrel}(A, Y, \mathcal{G}) = \emptyset$. Let $\mathbf{W} \equiv \mathrm{de}^c_{\mathcal{G}}(A)$ and $\mathbf{O} \equiv \mathbf{O}(A, Y, \mathcal{G})$. Write $\mathbf{W} \equiv (W_1, \ldots, W_J)$ and $\mathbf{O} \equiv (O_1, \ldots, O_T)$ in topological order relative to $\mathcal{G}$. Then, under $\mathcal{M}(\mathcal{G})$, the law of Y given $\mathbf{W}$ is the same as the law of Y given $(A, \mathbf{O})$ and the law of Y given $(A, \mathbf{O})$ is unrestricted. In particular, the conditional expectation $E(Y|\mathbf{W}) = E(Y|A, \mathbf{O})$ is unrestricted. Furthermore, the law of Y given $(A, \mathbf{O})$ and the law of $\mathbf{W} \cup \{A\}$ are variation independent.*

Proof [Proof of Lemma 32] That the law of Y given $\mathbf{W}$ is the same as the law of Y given $A, \mathbf{O}$ under any $P \in \mathcal{M}(\mathcal{G})$ follows because $Y \perp\!\!\!\perp_{\mathcal{G}} \mathbf{W} \setminus (A, \mathbf{O}) \,|\, (A, \mathbf{O})$.

Next, assume $\mathrm{de}_{\mathcal{G}}(A) \setminus \{Y\} \neq \emptyset$. Let $\mathcal{G}' = \mathcal{G}_{\mathbf{V} \setminus \mathrm{an}^c_{\mathcal{G}}(A,Y)}$ and $\mathbf{V}' = \mathbf{V} \setminus \mathrm{an}^c_{\mathcal{G}}(A, Y)$. Since $\mathbf{V} \setminus \mathrm{an}^c_{\mathcal{G}}(A, Y)$ is ancestral, $\mathcal{M}(\mathcal{G}') = \mathcal{M}(\mathcal{G}, \mathbf{V}')$ (see Proposition 1 from Evans (2016)). Let $\mathbf{M} \equiv (M_1, \ldots, M_K) \equiv \mathrm{de}_{\mathcal{G}}(A) \setminus \{Y\}$ be topologically ordered relative to $\mathcal{G}'$. Now, define $\mathcal{G}_K \equiv \tau(\mathcal{G}', M_K)$ and recursively for $k = K-1, K-2, \ldots, 1$ define $\mathcal{G}_k \equiv \tau(\mathcal{G}_{k+1}, M_k)$. Now,

since $\mathrm{ch}_{\mathcal{G}'}(M_K) = \{Y\}$, by Lemma 3 of Evans (2016), $\mathcal{M}(\mathcal{G}_K) = \mathcal{M}(\mathcal{G}', \mathbf{V}' \setminus \{M_K\})$. Furthermore, $\mathrm{pa}_{\mathcal{G}_K}(Y) = \mathrm{pa}_{\mathcal{G}'}(Y) \cup \mathrm{pa}_{\mathcal{G}'}(M_K)$.Likewise, since for $k = K-1, K-2, \ldots, 1$, $\mathrm{ch}_{\mathcal{G}_{k+1}}(M_k) = \{Y\}$, then we can recursively show that for $k = K-1, K-2, \ldots, 1$, $\mathcal{M}(\mathcal{G}_k) = \mathcal{M}(\mathcal{G}_{k+1}, \mathbf{V}' \setminus \{M_K, M_{K-1}, \ldots, M_k\})$ and $\mathrm{pa}_{\mathcal{G}_k}(Y) = \mathrm{pa}_{\mathcal{G}'}(Y) \cup \left[\cup_{l=k}^{K} \mathrm{pa}_{\mathcal{G}'}(M_l)\right]$. In particular, $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2, \mathbf{V}' \backslash \mathbf{M})$ and $\mathrm{pa}_{\mathcal{G}_1}(Y) = \mathrm{pa}_{\mathcal{G}}(Y) \cup \left[\cup_{l=1}^{K} \mathrm{pa}_{\mathcal{G}}(M_l)\right]$. Applying repeatedly the property (46) we arrive at $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}, \mathbf{V}' \backslash \mathbf{M})$. But $(A, \mathbf{O}) = \mathrm{pa}_{\mathcal{G}_1}(Y)$ and in $\mathcal{M}(\mathcal{G}_1)$ the law of Y given $\mathrm{pa}_{\mathcal{G}_1}(Y)$ is unrestricted. This implies that the law of Y given $(A, \mathbf{O})$ is unrestricted under $\mathcal{M}(\mathcal{G}_1)$. Then, $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}, \mathbf{V}' \backslash \mathbf{M})$ implies that the law of Y given $(A, \mathbf{O})$ is unrestricted under $\mathcal{M}(\mathcal{G})$. Finally, in model $\mathcal{M}(\mathcal{G}_1)$ (and consequently in model $\mathcal{M}(\mathcal{G})$) the law of $\mathrm{de}^c_{\mathcal{G}}(A) \cup \{A\}$ and the law of Y given $\mathrm{pa}_{\mathcal{G}_1}(Y)$ are variation independent, and therefore so are the laws of $\mathrm{de}^c_{\mathcal{G}}(A) \cup \{A\}$ and of Y given $(A, \mathbf{O})$ under model $\mathcal{M}(\mathcal{G})$.

If $\mathrm{de}_{\mathcal{G}}(A) \setminus \{Y\} = \emptyset$ then $(A, \mathbf{O}) = \mathrm{pa}_{\mathcal{G}}(Y)$ and the result follows immediately arguing as above. ■

Lemma 33 *Assume that $\mathcal{G}$ is a DAG with vertex set $\mathbf{V}$ and A and Y are two distinct vertices in $\mathcal{G}$ such that $A \in \mathrm{an}_{\mathcal{G}}(Y)$. Let $\mathbf{W} \equiv \mathrm{de}^c_{\mathcal{G}}(A)$ and $\mathbf{O} \equiv \mathbf{O}(A, Y, \mathcal{G})$. Write $\mathbf{W} \equiv (W_1, \ldots, W_J)$ and $\mathbf{O} \equiv (O_1, \ldots, O_T)$ in topological order relative to $\mathcal{G}$. Assume $\mathbf{O} \backslash O_T \subset \mathrm{pa}_{\mathcal{G}}(O_T)$ and $\mathrm{irrel}(A, Y, \mathcal{G}) = \emptyset$. Let $\mathbf{I}_j \equiv \left[\mathrm{pa}_{\mathcal{G}}(W_j) \cup \{W_j\}\right] \cap \mathrm{pa}_{\mathcal{G}}(W_{j+1})$. Assume that $J > 1$ and for some $j \in \{1, \ldots, J-1\}$ it holds that for $k = j+1, \ldots, J-1$,*

$$\mathbf{O} \backslash \mathbf{I}_k \perp\!\!\!\perp_{\mathcal{G}} \left[\mathrm{pa}_{\mathcal{G}}(W_k) \cup W_k\right] \triangle \mathrm{pa}_{\mathcal{G}}(W_{k+1}) \,|\mathbf{I}_k \tag{88}$$

and

$$\mathbf{O} \backslash \mathbf{I}_j \not\perp\!\!\!\perp_{\mathcal{G}} \left[\mathrm{pa}_{\mathcal{G}}(W_j) \cup W_j\right] \triangle \mathrm{pa}_{\mathcal{G}}(W_{j+1}) \,|\mathbf{I}_j \tag{89}$$

where the assertion (88) is inexistant if $j = J-1$. Then there exists $P^* \in \mathcal{M}(\mathcal{G})$ such that

$$E_{P^*}\left[b_a(\mathbf{O}; P^*) \,|W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right] - E_{P^*}\left[b_a(\mathbf{O}; P^*) \,|\, \mathrm{pa}_{\mathcal{G}}(W_{j+1})\right] \tag{90}$$

is a non-constant function of W_j.

Proof [Proof of Lemma 33]

First we show that if (88) holds for some $k \in \{1, ..., J-1\}$, then for such k it holds that $W_k \in \mathrm{pa}_{\mathcal{G}}(W_{k+1})$. Assume for the sake of contradiction that $W_k \notin \mathrm{pa}_{\mathcal{G}}(W_{k+1})$. Since $\mathrm{irrel}(A, Y, \mathcal{G}) = \emptyset$, there exists a directed path between W_k and Y that does not intersect A. Such a path must intersect $\mathbf{O}$. Since $\mathbf{O} \subset \mathrm{an}_{\mathcal{G}}(O_T)$ we conclude that $W_k \in \mathrm{an}_{\mathcal{G}}(O_T)$. Note also that $\mathbf{I}_k \cap \mathrm{de}_{\mathcal{G}}(W_k) = \emptyset$. Then

$$O_T \not\perp\!\!\!\perp_{\mathcal{G}} W_k \mid \mathbf{I}_k. \tag{91}$$

Now $O_T \in \mathbf{O} \setminus \mathbf{I}_k$ because $O_T = W_J$. Since $W_k \notin \mathrm{pa}_{\mathcal{G}}(W_{k+1})$ then $W_k \in \left[\mathrm{pa}_{\mathcal{G}}(W_k) \cup W_k\right] \triangle \mathrm{pa}_{\mathcal{G}}(W_{k+1})$ which together with (91) implies $\mathbf{O} \backslash \mathbf{I}_k \not\perp\!\!\!\perp_{\mathcal{G}} \left[\mathrm{pa}_{\mathcal{G}}(W_k) \cup W_k\right] \triangle \mathrm{pa}_{\mathcal{G}}(W_{k+1}) \,|\mathbf{I}_k$. This d-connection contradicts (88), thus proving that $W_k \in \mathrm{pa}_{\mathcal{G}}(W_{k+1})$.

We will show that for some $P^* \in \mathcal{M}(\mathcal{G})$, (90) is a non-constant function of W_j by considering separately the cases $W_j \notin \mathrm{pa}_{\mathcal{G}}(W_{j+1})$ and $W_j \in \mathrm{pa}_{\mathcal{G}}(W_{j+1})$.

Suppose first that $W_j \notin \text{pa}_{\mathcal{G}}(W_{j+1})$. Then, since $E_P\left[b_a(\mathbf{O};P)\,|\text{pa}_{\mathcal{G}}(W_{j+1})\right]$ does not depend on W_j for all $P \in \mathcal{M}(\mathcal{G})$, it suffices to prove that there exists $P^* \in \mathcal{M}(\mathcal{G})$ such that $E_{P^*}\left[b_a(\mathbf{O};P^*)\,|W_j, \text{pa}_{\mathcal{G}}(W_j)\right]$ is a non-constant function of W_j. To show this, first note that since $\text{irrel}(A,Y,\mathcal{G}) = \emptyset$ there exists a directed path between W_j and Y that does not intersect A. Such a path must intersect $\mathbf{O}$. Since $\mathbf{O}\backslash O_T \subset \text{pa}_{\mathcal{G}}(O_T)$, then $W_j \in \text{an}_{\mathcal{G}}(O_T)$. Consequently,

$$W_j \not\perp\!\!\!\perp_{\mathcal{G}} O_T \mid \text{pa}_{\mathcal{G}}(W_j). \tag{92}$$

Thus, there exists a law $P^* \in \mathcal{M}(\mathcal{G})$ such that under P^*, $W_j \not\perp\!\!\!\perp O_T \mid \text{pa}_{\mathcal{G}}(W_j)$. In particular, there exists a function $b^*(O_T)$ such that $E_{P^*}\left[b^*(O_T)\,|W_j, \text{pa}_{\mathcal{G}}(W_j)\right]$ is a non-constant function of W_j. Lemma 32 implies that we can choose the law P^* so that $b_a(\mathbf{O};P^*) = b^*(O_T)$ showing that for such law P^*, $E_{P^*}\left[b(\mathbf{O};P^*)\,|W_j, \text{pa}_{\mathcal{G}}(W_j)\right] = E_{P^*}\left[b^*(O_T)\,|W_j, \text{pa}_{\mathcal{G}}(W_j)\right]$ is a non-constant function of W_j, and consequently the difference (90) depends on W_j.

Suppose next that $W_j \in \text{pa}_{\mathcal{G}}(W_{j+1})$. For each $i = 1, \ldots, J-1$, define $\mathbf{O}_f^i \equiv (W_{i+1}, \ldots, W_J) \cap \mathbf{O}$ and $\mathbf{O}_p^i \equiv \mathbf{O}\backslash\mathbf{O}_f^i$. The vertex set $\mathbf{O}_f^i$ is not empty because $W_J = O_T$. Write $\mathbf{O}_f^i = \left(O_{f,1}^i, \ldots, O_{f,h}^i\right)$ and $\mathbf{O}_f^i = \left(O_{p,1}^i, \ldots, O_{p,m}^i\right)$ in topological order relative to $\mathcal{G}$. The validity of (89) is equivalent to the existence of $O \in \mathbf{O}\backslash\mathbf{I}_j$ and of $W \in \left[\text{pa}_{\mathcal{G}}(W_j) \cup W_j\right] \triangle \text{pa}_{\mathcal{G}}(W_{j+1})$ such that

$$O \not\perp\!\!\!\perp_{\mathcal{G}} W | \mathbf{I}_j \tag{93}$$

We will next show that if W is in $\text{pa}_{\mathcal{G}}(W_{j+1}) \setminus \left[W_j \cup \text{pa}_{\mathcal{G}}(W_j)\right]$, then (93) holds for $O = O_{f,1}^1$. So we will consider separately the following three cases

Case	O in	W in
1	$\left\{O_{f,1}^1\right\}$	$\text{pa}_{\mathcal{G}}(W_{j+1}) \setminus \left[W_j \cup \text{pa}_{\mathcal{G}}(W_j)\right]$
2	$\mathbf{O}_f^j$	$\text{pa}_{\mathcal{G}}(W_j) \setminus \text{pa}_{\mathcal{G}}(W_{j+1})$
3	$\mathbf{O}_p^j \backslash \mathbf{I}_j$	$\text{pa}_{\mathcal{G}}(W_j) \setminus \text{pa}_{\mathcal{G}}(W_{j+1})$

Notice that $\mathbf{O}_f^j \subset \mathbf{O}\backslash\mathbf{I}_j$ and that $\text{pa}_{\mathcal{G}}(W_j) \setminus \text{pa}_{\mathcal{G}}(W_{j+1}) = \left[W_j \cup \text{pa}_{\mathcal{G}}(W_j)\right] \setminus \text{pa}_{\mathcal{G}}(W_{j+1})$ because we have assumed that $W_j \in \text{pa}_{\mathcal{G}}(W_{j+1})$.

In the subsequent analysis we will use the fact that W_j and W_{j+1} belong to $\text{an}_{\mathcal{G}}\left(O_{f,1}^j\right)$. To see why this is true, first note that if $j = J-1$, then $W_j = W_{J-1} \in \text{pa}_{\mathcal{G}}(W_{j+1}) = \text{pa}_{\mathcal{G}}(W_J)$
$= \text{pa}_{\mathcal{G}}(O_T)$ by assumption. On the other hand, $\mathbf{O}_f^{J-1} = O_{f,1}^{J-1} = O_T$. Then, $W_j = W_{J-1}$ and $W_{j+1} = W_J$ belong to $\text{an}_{\mathcal{G}}(O_T) = \text{an}_{\mathcal{G}}\left(O_{f,1}^{J-1}\right) = \text{an}_{\mathcal{G}}\left(O_{f,1}^j\right)$. If $j < J-1$, then W_j and W_{j+1} also belong to $\text{an}_{\mathcal{G}}\left(O_{f,1}^j\right)$ because we have already shown that $W_k \in \text{pa}_{\mathcal{G}}(W_{k+1})$ for all $k = j+1, \ldots, J-1$ and by definition $O_{f,1}^j \in \{W_{j+1}, ..., W_J\}$.

Consider the case (1). The vertex W belongs to $\mathbf{O}\left(W_j, O_{f,1}^j, \mathcal{G}\right)$ by virtue of being an element of the parent set of the child W_{j+1} of W_j and the facts that (i) $W \in \text{de}_{\mathcal{G}}^c(W_j)$ because it belongs to $\text{pa}_{\mathcal{G}}(W_{j+1}) \setminus \{W_j\}$ and (ii) W_j and W_{j+1} belong to $\text{an}_{\mathcal{G}}\left(O_{f,1}^j\right)$. Note that this implies that $W \in \text{an}_{\mathcal{G}}\left(O_{f,1}^j\right)$ and consequently that (93) holds with $O = O_{f,1}^j$ because the path $W \to W_{j+1} \to W_{j+2} \to ... \to O_{f,1}^j$ is open given $\mathbf{I}_j$ since $\mathbf{I}_j$ does not

include any of the nodes in the set $\left\{W_{j+1}, W_{j+2}, ..., O^j_{f,1}\right\}$. Now, Lemma 32 implies that $E_P\left[O^j_{f,1}|W_j, \mathbf{O}\left(W_j, O^j_{f,1}; \mathcal{G}\right)\right]$ is unrestricted in model $\mathcal{M}(\mathcal{G})$ so we can choose P^* such that $E_{P^*}\left[O^j_{f,1}|W_j, \mathbf{O}\left(W_j, O^j_{f,1}; \mathcal{G}\right)\right] = W_j W$. We can also choose such P^* so as to also satisfy that $b_a(\mathbf{O}; P^*) = O^j_{f,1}$ where $\mathbf{O} \equiv \mathbf{O}(A, Y; \mathcal{G})$. This can be done because, as established in Lemma 32, the conditional law of Y given $(A, \mathbf{O})$ is variation independent with the joint law of $\mathbf{W}$, and in particular, with the joint law of the subvector $\left(O^j_{f,1}, W_j, \mathbf{O}_{W_j}\right)$ of $\mathbf{W}$. Then, $E_{P^*}\left[b_a(\mathbf{O}; P^*)|\mathrm{pa}_{\mathcal{G}}(W_{j+1})\right] = E_{P^*}\left[O^j_{f,1}|\mathrm{pa}_{\mathcal{G}}(W_{j+1})\right] =$
$E_{P^*}\left\{E_P\left[O^j_{f,1}|W_j, \mathbf{O}_{W_j}, \mathrm{pa}_{\mathcal{G}}(W_{j+1})\right]\middle|\mathrm{pa}_{\mathcal{G}}(W_{j+1})\right\} =$
$E_{P^*}\left\{E_{P^*}\left[O^j_{f,1}|W_j, \mathbf{O}_{W_j}\right]\middle|\mathrm{pa}_{\mathcal{G}}(W_{j+1})\right\} = W_j W$. Consequently,
$E_{P^*}\left[b_a(\mathbf{O}; P^*)|W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right] - E_{P^*}\left[b_a(\mathbf{O}; P)|\mathrm{pa}_{\mathcal{G}}(W_{j+1})\right] =$
$E_{P^*}\left[b_a(\mathbf{O}; P^*)|W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right] - W_j W$ which depends on W_j because $W \in \mathrm{pa}_{\mathcal{G}}(W_{j+1}) \setminus \left[W_j \cup \mathrm{pa}_{\mathcal{G}}(W_j)\right]$.

Consider now case (2). Let $W \in \mathrm{pa}_{\mathcal{G}}(W_j) \setminus \mathrm{pa}_{\mathcal{G}}(W_{j+1})$ and $O^j_{f,l} \in \mathbf{O} \setminus \mathbf{I}_j$ such that $O^j_{f,l} \not\perp\!\!\!\perp_{\mathcal{G}} W|\mathbf{I}_j$.Let τ denote a path between $O^j_{f,l}$ and W that is open given $\mathbf{I}_j$. In τ the edge with one endpoint equal to $O^j_{f,l}$ must point into $O^j_{f,l}$. Suppose this was not the case, then τ would intersect a collider, say C, that is a descendant of $O^j_{f,l}$. However, by the definitions of $\mathbf{I}_j$ and $O^j_{f,l}$ we know that $\mathbf{I}_j \cap \mathrm{de}_{\mathcal{G}}\left(O^j_{f,l}\right) = \emptyset$. Consequently C cannot have a descendant in $\mathbf{I}_j$. So, the path τ would be blocked at C given $\mathbf{I}_j$ contradicting that τ is open given $\mathbf{I}_j$. Because the path τ is open, then it must intersect an element of the set $\mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right)$, and consequently, $\mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right) \not\perp\!\!\!\perp_{\mathcal{G}} W|\mathbf{I}_j$. Now, $W \in \mathrm{pa}(W_j) \setminus \mathbf{I}_j$ because $\left[\{W_j\} \cup \mathrm{pa}_{\mathcal{G}}(W_j)\right] \setminus \mathrm{pa}_{\mathcal{G}}(W_{j+1}) = \mathrm{pa}_{\mathcal{G}}(W_j) \setminus \mathbf{I}_j$ since we have assumed that $W_j \in \mathrm{pa}_{\mathcal{G}}(W_{j+1})$. We then conclude that

$$\mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right) \not\perp\!\!\!\perp_{\mathcal{G}} \left[\mathrm{pa}_{\mathcal{G}}(W_j) \setminus \mathbf{I}_j\right]|\mathbf{I}_j. \tag{94}$$

So, there exists $P^* \in \mathcal{M}(\mathcal{G})$ such that

$$\mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right) \not\perp\!\!\!\perp \left[\mathrm{pa}_{\mathcal{G}}(W_j) \setminus \mathbf{I}_j\right]|\mathbf{I}_j \text{ under } P^*. \tag{95}$$

Now, (95) implies that there exists $h^*\left[\mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right)\right]$ such that
$E_{P^*}\left\{h^*\left[\mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right)\right]\middle|\left[\mathrm{pa}(W_j) \setminus \mathbf{I}_j\right], \mathbf{I}_j\right\}$ is a non-constant function of $\mathrm{pa}(W_j) \setminus \mathbf{I}_j$. Then, since $\left[\mathrm{pa}(W_j) \setminus \mathbf{I}_j\right] \cup \mathbf{I}_j = \mathrm{pa}_{\mathcal{G}}(W_j) \cup W_j$ we conclude that
$E_{P^*}\left\{h^*\left[\mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right)\right]\middle|\mathrm{pa}_{\mathcal{G}}(W_j), W_j\right\}$is a non-constant function of $\mathrm{pa}(W_j) \setminus \mathbf{I}_j$. Furthermore, by the Local Markov property,
$E_{P^*}\left\{h^*\left[\mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right)\right]\middle|\mathrm{pa}_{\mathcal{G}}(W_j), W_j\right\}$ does not depend on W_j. So, we conclude that
$E_{P^*}\left\{h^*\left[\mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right)\right]\middle|\mathrm{pa}_{\mathcal{G}}(W_j), W_j\right\} = g\left[\mathrm{pa}_{\mathcal{G}}(W_j)\right]$ where $g\left[\mathrm{pa}_{\mathcal{G}}(W_j)\right]$ is a non-constant function of $\mathrm{pa}(W_j) \setminus \mathbf{I}_j$.

Now, by the variation independence of the conditional law of Y given $(A, \mathbf{O})$ with the joint law of $\mathbf{W}$, which holds as established in Lemma 32, we can take P^* to also satisfy $b_a(\mathbf{O}; P^*) = O^j_{f,l}$. Furthermore, we can take P^* to additionally satisfy that $E_{P^*}\left[O^j_{f,l} \middle| W_j, \mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right)\right] = W_j h^*\left[\mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right)\right]$ because, again by Lemma 32, the conditional law of $O^j_{f,l}$ given $W_j, \mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right)$ is variation independent with the law of $\mathrm{de}^c_{\mathcal{G}}(W_j) \cup W_j$, and in particular, with the joint law of law of $\mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right)$ and $\left[W_j \cup \mathrm{pa}_{\mathcal{G}}(W_j)\right]$. For such P^* we then have

$$E_{P^*}\left[b_a(\mathbf{O}; P^*) | W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right] = E_{P^*}\left[O^j_{f,l} | W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right] =$$
$$E_{P^*}\left[E_{P^*}\left[O^j_{f,l} | W_j, \mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right)\right] \middle| W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right] =$$
$$W_j E_{P^*}\left[h^*\left[\mathbf{O}\left(W_j, O^j_{f,l}, \mathcal{G}\right)\right] \middle| W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right] = W_j g\left[\mathrm{pa}_{\mathcal{G}}(W_j)\right].$$

Then, $E_{P^*}\left[b_a(\mathbf{O}; P^*) | W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right] - E_{P^*}\left[b_a(\mathbf{O}; P^*) \,|\, \mathrm{pa}_{\mathcal{G}}(W_{j+1})\right] = W_j g\left[\mathrm{pa}_{\mathcal{G}}(W_j)\right] - E_{P^*}\left[b_a(\mathbf{O}; P^*) \,|\, \mathrm{pa}_{\mathcal{G}}(W_{j+1})\right]$ is a non-constant function of W_j because $g\left[\mathrm{pa}_{\mathcal{G}}(W_j)\right]$ is a non-constant function of $\mathrm{pa}(W_j) \setminus \mathbf{I}_j$ and $\left[\mathrm{pa}_{\mathcal{G}}(W_j)\right] \setminus \mathbf{I}_j \cap \mathrm{pa}_{\mathcal{G}}(W_{j+1}) = \emptyset$.

Finally, consider case (3). Let $W \in \mathrm{pa}_{\mathcal{G}}(W_j) \setminus \mathrm{pa}_{\mathcal{G}}(W_{j+1})$ and $O^j_{p,l} \in \mathbf{O}^j_p \setminus \mathbf{I}_j$ such that $O^j_{p,l} \not\perp\!\!\!\perp_{\mathcal{G}} W | \mathbf{I}_j$.We then have that $O^j_{p,l} \not\perp\!\!\!\perp_{\mathcal{G}} \left[\mathrm{pa}_{\mathcal{G}}(W_j) \setminus \mathbf{I}_j\right] | \mathbf{I}_j$, which then implies that there exists $P^* \in \mathcal{M}(\mathcal{G})$ such that $O^j_{p,l} \not\perp\!\!\!\perp \left[\mathrm{pa}_{\mathcal{G}}(W_j) \setminus \mathbf{I}_j\right] | \mathbf{I}_j$ under P^*. The latter implies that there exists $h^*\left(O^j_{p,l}\right)$ such that $E_{P^*}\left\{h^*\left(O^j_{p,l}\right) \middle| \left[\mathrm{pa}(W_j) \setminus \mathbf{I}_j\right], \mathbf{I}_j\right\}$ is a non-constant function of $\mathrm{pa}(W_j) \setminus \mathbf{I}_j$. Then, since $\left[\mathrm{pa}(W_j) \setminus \mathbf{I}_j\right] \cup \mathbf{I}_j = \mathrm{pa}_{\mathcal{G}}(W_j) \cup W_j$ we conclude that $E_{P^*}\left\{h^*\left(O^j_{p,l}\right) \middle| \mathrm{pa}_{\mathcal{G}}(W_j), W_j\right\}$ is a non-constant function of $\mathrm{pa}(W_j) \setminus \mathbf{I}_j$. Furthermore, by the Local Markov property, $E_{P^*}\left\{h^*\left(O^j_{p,l}\right) \middle| \mathrm{pa}_{\mathcal{G}}(W_j), W_j\right\}$ does not depend on W_j. So, we conclude that $E_{P^*}\left\{h^*\left(O^j_{p,l}\right) \middle| \mathrm{pa}_{\mathcal{G}}(W_j), W_j\right\} = g\left[\mathrm{pa}_{\mathcal{G}}(W_j)\right]$ where $g\left[\mathrm{pa}_{\mathcal{G}}(W_j)\right]$ is a non-constant function of $\mathrm{pa}(W_j) \setminus \mathbf{I}_j$. By Lemma 32 we can take P^* to also satisfy that $b(\mathbf{O}; P^*) = h^*\left(O^j_{p,l}\right) O^j_{f,1}$ and $E_{P^*}\left[O^j_{f,1} \middle| W_j, \mathbf{O}\left(W_j, O^j_{f,1}, \mathcal{G}\right)\right] = W_j$. Then $E_{P^*}\left[b_a(\mathbf{O}; P^*) | W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right] = E_{P^*}\left\{h^*\left(O^j_{p,l}\right) O^j_{f,1} \middle| W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right\} =$
$E_{P^*}\left\{h^*\left(O^j_{p,l}\right) E_P\left[O^j_{f,1} \middle| W_j, \mathbf{O}\left(W_j, O^j_{f,1}, \mathcal{G}\right)\right] \middle| W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right\} =$
$W_j E_{P^*}\left\{h^*\left(O^j_{p,l}\right) \middle| W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right\} = W_j g\left[\mathrm{pa}_{\mathcal{G}}(W_j)\right]$. Consequently,
$E_{P^*}\left[b_a(\mathbf{O}; P^*) | W_j, \mathrm{pa}_{\mathcal{G}}(W_j)\right] - E_{P^*}\left[b_a(\mathbf{O}; P^*) | \mathrm{pa}_{\mathcal{G}}(W_{j+1})\right] = W_j g\left[\mathrm{pa}_{\mathcal{G}}(W_j)\right] -$
$E_P\left[b_a(\mathbf{O}; P^*) | \mathrm{pa}_{\mathcal{G}}(W_{j+1})\right]$ depends on W_j. This concludes the proof of the lemma. ■

Lemma 34 *Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$ and let A and Y be two distinct vertices in $\mathbf{V}$. Suppose that* $\mathrm{irrel}(A, Y; \mathcal{G}) = \emptyset$. *Suppose* $\mathbf{M} \equiv \mathrm{de}_{\mathcal{G}}(A) \setminus \{A, Y\} \neq \emptyset$ *and let $(M_1, \ldots, M_K)$ be the elements of $\mathbf{M}$ sorted topologically. Let $M_0 \equiv A$ and $M_{K+1} \equiv Y$. Let $\mathbf{O} \equiv \mathbf{O}(A, Y, \mathcal{G})$. Let $\mathbf{O}_{min}$ be the smallest among the subsets $\mathbf{O}_{sub}$ of $\mathbf{O}$ such that $A \perp\!\!\!\perp_{\mathcal{G}} (\mathbf{O} \setminus \mathbf{O}_{sub}) | \mathbf{O}_{sub}$.*

1. *$E_P\left[T_{P,a,\mathcal{G}} | Y, \mathrm{pa}_{\mathcal{G}}(Y)\right] = T_{P,a,\mathcal{G}}$ for all $P \in \mathcal{M}(\mathcal{G})$ if and only if $\{A\} \cup \mathbf{O}_{\min} \subseteq \mathrm{pa}_{\mathcal{G}}(Y)$.*

2. *Suppose* $\{A\} \cup \mathbf{O}_{\min} \subseteq \mathrm{pa}_{\mathcal{G}}(Y)$. *If* $\mathrm{pa}_{\mathcal{G}}(Y) \setminus \{M_K\} \not\subset \mathrm{pa}_{\mathcal{G}}(M_K)$ *then there exists* $P \in \mathcal{M}(\mathcal{G})$ *such that* $E_P\left[T_{P,a,\mathcal{G}}|M_K, \mathrm{pa}_{\mathcal{G}}(M_K)\right] - E_P\left[T_{P,a,\mathcal{G}} \mid \mathrm{pa}_{\mathcal{G}}(Y)\right]$ *is a non-constant function of* M_K.

3. *Suppose* $\{A\} \cup \mathbf{O}_{\min} \subseteq \mathrm{pa}_{\mathcal{G}}(Y), \mathrm{pa}_{\mathcal{G}}(Y) \setminus \{M_K\} \subset \mathrm{pa}_{\mathcal{G}}(M_K)$ *and there exists* $j \geq 1$ *such that for all* $k = K-1, \ldots, j+1, \mathrm{pa}_{\mathcal{G}}(M_{k+1}) \setminus \{M_k\} \subset \mathrm{pa}_{\mathcal{G}}(M_k)$ *but* $\mathrm{pa}_{\mathcal{G}}(M_{j+1}) \setminus \{M_j\} \not\subset \mathrm{pa}_{\mathcal{G}}(M_j)$. *Then, there exists* $P \in \mathcal{M}(\mathcal{G})$ *such that* $E_P\left[T_{P,a,\mathcal{G}}|M_j, \mathrm{pa}_{\mathcal{G}}(M_j)\right] - E_P\left[T_{P,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(M_{j+1})\right]$ *is a non-constant of function of* M_j.

Proof [Proof of Lemma 34]
1) If $\{A\} \cup \mathbf{O}_{\min} \subseteq \mathrm{pa}_{\mathcal{G}}(Y)$, then $E_P\left[T_{P,a,\mathcal{G}}|Y, \mathrm{pa}_{\mathcal{G}}(Y)\right] = T_{P,a,\mathcal{G}}$ for all $P \in \mathcal{M}(\mathcal{G})$ holds trivially by the definition of $T_{P,a,\mathcal{G}}$.

Now suppose that $A \notin \mathrm{pa}_{\mathcal{G}}(Y)$ or $\mathbf{O}_{\min} \not\subset \mathrm{pa}_{\mathcal{G}}(Y)$. If $A \notin \mathrm{pa}_{\mathcal{G}}(Y)$ then $E_P\left[T_{P,a,\mathcal{G}}|Y, \mathrm{pa}_{\mathcal{G}}(Y)\right]$ is not a function of A and consequently, it cannot be equal to $I_a(A)Y/\pi_a(\mathbf{O}; P)$. Next, suppose $\mathbf{O}_{min} \not\subseteq \mathrm{pa}_{\mathcal{G}}(Y)$ because for some $O_j \in \mathbf{O}_{min}, O_j \notin \mathrm{pa}_{\mathcal{G}}(Y)$. Now, because $\mathbf{O}_{min}$ is the smallest among the subsets $\mathbf{O}_{sub}$ of $\mathbf{O}$ such that $A \perp\!\!\!\perp_{\mathcal{G}} (\mathbf{O} \setminus \mathbf{O}_{sub}) | \mathbf{O}_{sub}$, then there exists a law $P^* \in \mathcal{M}(\mathcal{G})$ such that $I_a(A)Y/\pi_a(\mathbf{O}_{min}; P^*)$ is a non-constant function of O_j. For such P^*, $E_{P^*}\left[T_{P^*,a,\mathcal{G}}|Y, \mathrm{pa}_{\mathcal{G}}(Y)\right]$ cannot be equal to $I_a(A)Y/\pi_a(\mathbf{O}_{min}; P^*)$.

2) Suppose that $\{A\} \cup \mathbf{O}_{\min} \subset \mathrm{pa}_{\mathcal{G}}(Y)$ but $\mathrm{pa}_{\mathcal{G}}(Y) \setminus \{M_K\} \not\subset \mathrm{pa}_{\mathcal{G}}(M_K)$. Let $M^* \in \mathrm{pa}_{\mathcal{G}}(Y) \setminus \left\{M_K \cup \mathrm{pa}_{\mathcal{G}}(M_K)\right\}$. Since M_K is the last element in the topological order of $\mathbf{M}$ and the assumptions that $\mathrm{irrel}(A, Y, \mathcal{G})$, $M_K \in \mathrm{pa}_{\mathcal{G}}(Y)$. Then there exists $P^* \in \mathcal{M}(G)$ be such that $E_{P^*}\left[Y|\,\mathrm{pa}_{\mathcal{G}}(Y)\right] = M^*M_K$. For such P^*, $E_{P^*}\left[T_{P^*,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(Y)\right] = \frac{A}{\pi(\mathbf{O}_{\min}; P^*)}M^*M_K$. Furthermore,

$$
\begin{aligned}
& E_{P^*}\left[T_{P^*,a,\mathcal{G}}|M_K, \mathrm{pa}_{\mathcal{G}}(M_K)\right] \\
= \; & E_{P^*}\left[\frac{I_a(A)}{\pi_a(\mathbf{O}_{min}; P^*)} E_{P^*}\left[Y|A, \mathbf{O}_{min}, M_K, \mathrm{pa}_{\mathcal{G}}(M_K), \mathrm{pa}_{\mathcal{G}}(Y)\right] \middle| M_K, \mathrm{pa}_{\mathcal{G}}(M_K)\right] \\
= \; & E_{P^*}\left[\frac{I_a(A)}{\pi_a(\mathbf{O}_{min}; P^*)} E_{P^*}\left[Y|\mathrm{pa}_{\mathcal{G}}(Y)\right] \middle| M_K, \mathrm{pa}_{\mathcal{G}}(M_K)\right] \\
= \; & E_{P^*}\left[\frac{I_a(A)}{\pi_a(\mathbf{O}_{min}; P^*)} M^*M_K \middle| M_K, \mathrm{pa}_{\mathcal{G}}(M_K)\right] \\
= \; & M_K E_{P^*}\left[\frac{I_a(A)}{\pi_a(\mathbf{O}_{min}; P^*)} M^* \middle| M_K, \mathrm{pa}_{\mathcal{G}}(M_K)\right].
\end{aligned}
$$

Then,

$$
\begin{aligned}
& E_{P^*}\left[T_{P^*,a,\mathcal{G}}|M_K, \mathrm{pa}_{\mathcal{G}}(M_K)\right] - E_{P^*}\left[T_{P^*,a,\mathcal{G}}|\,\mathrm{pa}_{\mathcal{G}}(Y)\right] \\
= \; & M_K\left\{E_{P^*}\left[\frac{I_a(A)}{\pi_a(\mathbf{O}_{min}; P^*)} M^* \middle| M_K, \mathrm{pa}_{\mathcal{G}}(M_K)\right] - \frac{A}{\pi(\mathbf{O}_{\min}; P^*)} M^*\right\}.
\end{aligned}
$$

The right hand side is a non-constant function of M_K because $M^* \notin \{M_K\} \cup \mathrm{pa}_{\mathcal{G}}(M_K)$.

3) Suppose that $\{A\} \cup \mathbf{O}_{\min} \subset \mathrm{pa}_{\mathcal{G}}(Y)$ and $\mathrm{pa}_{\mathcal{G}}(Y) \setminus \{M_K\} \subset \mathrm{pa}_{\mathcal{G}}(M_K)$ and that $\mathrm{pa}_{\mathcal{G}}(M_{k+1}) \setminus \{M_k\} \subset \mathrm{pa}_{\mathcal{G}}(M_k)$ for all $k = K-1, \ldots, j+1$, but $\mathrm{pa}_{\mathcal{G}}(M_{j+1}) \setminus \{M_j\} \not\subset$

$\mathrm{pa}_{\mathcal{G}}(M_j)$. Now $\mathrm{pa}_{\mathcal{G}}(M_{j+1}) \setminus \{M_j\} \not\subset \mathrm{pa}_{\mathcal{G}}(M_j)$ implies that there exists $M^{**} \in \mathrm{pa}_{\mathcal{G}}(M_{j+1}) \setminus \left\{M_j \cup \mathrm{pa}_{\mathcal{G}}(M_j)\right\}$. On the other hand, by part (i) of Lemma 31 we know that $M_k \in \mathrm{pa}_{\mathcal{G}}(M_{k+1})$ for $k = j, \ldots, K$. Now, consider a law P^* such that $E_{P^*}\left[Y \mid \mathrm{pa}_{\mathcal{G}}(Y)\right] = M_K$ and $E_{P^*}\left[M_k \mid \mathrm{pa}_{\mathcal{G}}(M_k)\right] = M_{k-1}$ for all $k = j+2, \ldots, K$ and such that $E_{P^*}\left[M_{j+1} \mid \mathrm{pa}_{\mathcal{G}}(M_{j+1})\right] = M^{**}M_j$. Since $\{A\} \cup \mathbf{O}_{min} \subset \mathrm{pa}_{\mathcal{G}}(Y) \subset \{M_K\} \cup \mathrm{pa}_{\mathcal{G}}(M_K) \subset \cdots \subset \mathrm{pa}_{\mathcal{G}}(M_{j+1})$ then

$$
\begin{aligned}
&E_{P^*}\left[T_{P^*,a,\mathcal{G}} | \mathrm{pa}_{\mathcal{G}}(M_{j+1})\right] = \\
&\frac{I_a(A)}{\pi_a(\mathbf{O}_{min}; P^*)} E_{P^*}\left[E_{P^*}\left[Y | \mathrm{pa}_{\mathcal{G}}(Y), \mathrm{pa}_{\mathcal{G}}(M_{j+1})\right] \middle| \mathrm{pa}_{\mathcal{G}}(M_{j+1})\right] = \\
&\frac{I_a(A)}{\pi_a(\mathbf{O}_{min}; P^*)} E_{P^*}\left[E_{P^*}\left[Y | \mathrm{pa}_{\mathcal{G}}(Y)\right] \middle| \mathrm{pa}_{\mathcal{G}}(M_{j+1})\right] = \\
&\frac{I_a(A)}{\pi_a(\mathbf{O}_{min}; P^*)} E_{P^*}\left[M_K \mid \mathrm{pa}_{\mathcal{G}}(M_{j+1})\right] = \\
&\frac{I_a(A)}{\pi_a(\mathbf{O}_{min}; P^*)} E_{P^*}\left[E_{P^*}\left[M_K | \mathrm{pa}_{\mathcal{G}}(M_K), \mathrm{pa}_{\mathcal{G}}(M_{j+1})\right] \middle| \mathrm{pa}_{\mathcal{G}}(M_{j+1})\right] = \\
&\frac{I_a(A)}{\pi_a(\mathbf{O}_{min}; P^*)} E_{P^*}\left[E_{P^*}\left[M_K | \mathrm{pa}_{\mathcal{G}}(M_K)\right] \middle| \mathrm{pa}_{\mathcal{G}}(M_{j+1})\right] = \\
&\frac{I_a(A)}{\pi_a(\mathbf{O}_{min}; P^*)} E_{P^*}\left[M_{K-1} \middle| \mathrm{pa}_{\mathcal{G}}(M_{j+1})\right] = \\
&\cdots = \\
&\frac{I_a(A)}{\pi_a(\mathbf{O}_{min}; P^*)} E_{P^*}\left[M_{j+1} \middle| \mathrm{pa}_{\mathcal{G}}(M_{j+1})\right] = \\
&\frac{I_a(A)}{\pi_a(\mathbf{O}_{min}; P^*)} M^{**}M_j.
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
& E_{P^*}\left[T_{P^*,a,\mathcal{G}}|M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right] \\
= \; & E_{P^*}\left[\frac{I_a(A)}{\pi_a\left(\mathbf{O}_{min}; P^*\right)}\, E_{P^*}\left[Y|A, \mathbf{O}_{min}, \mathrm{pa}_{\mathcal{G}}\left(Y\right), M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right]\Big|\, M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right] \\
= \; & E_{P^*}\left[\frac{I_a(A)}{\pi_a\left(\mathbf{O}_{min}; P^*\right)}\, E_{P^*}\left[Y|\mathrm{pa}_{\mathcal{G}}\left(Y\right)\right]\Big|\, M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right] \\
= \; & E_{P^*}\left[\frac{I_a(A)}{\pi_a\left(\mathbf{O}_{min}; P^*\right)} M_K \mid M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right] \\
= \; & E_{P^*}\left[\frac{I_a(A)}{\pi_a\left(\mathbf{O}_{min}; P^*\right)}\, E_{P^*}\left[M_K|A, \mathbf{O}_{min}, \mathrm{pa}_{\mathcal{G}}\left(M_K\right), M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right]\Big|\, M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right] \\
= \; & E_{P^*}\left[\frac{I_a(A)}{\pi_a\left(\mathbf{O}_{min}; P^*\right)}\, E_{P^*}\left[M_K|\mathrm{pa}_{\mathcal{G}}\left(M_K\right)\right]\Big|\, M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right] \\
= \; & E_{P^*}\left[\frac{I_a(A)}{\pi_a\left(\mathbf{O}_{min}; P^*\right)}\, M_{K-1}|\, M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right] \\
= \; & \ldots \\
= \; & E_{P^*}\left[\frac{I_a(A)}{\pi_a\left(\mathbf{O}_{min}; P^*\right)}\, M_{j+1}|\, M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right] \\
= \; & E_{P^*}\left[\frac{I_a(A)}{\pi_a\left(\mathbf{O}_{min}; P^*\right)}\, E_{P^*}\left[M_{j+1} \mid A, \mathbf{O}_{min}, M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right), \mathrm{pa}_{\mathcal{G}}\left(M_{j+1}\right)\right]\Big|\, M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right] \\
= \; & E_{P^*}\left[\frac{I_a(A)}{\pi_a\left(\mathbf{O}_{min}; P^*\right)}\, E_{P^*}\left[M_{j+1} \mid \mathrm{pa}_{\mathcal{G}}\left(M_{j+1}\right)\right]\Big|\, M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right] \\
= \; & E_{P^*}\left[\frac{I_a(A)}{\pi_a\left(\mathbf{O}_{min}; P^*\right)}\, M^{**} M_j|\, M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right] \\
= \; & M_j E_{P^*}\left[\frac{I_a(A)}{\pi_a\left(\mathbf{O}_{min}; P^*\right)}\, M^{**}|\, M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right].
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
& E_{P^*}\left[T_{P^*,a,\mathcal{G}}|M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right] - E_{P^*}\left[T_{P^*,a,\mathcal{G}}|\mathrm{pa}_{\mathcal{G}}\left(M_{j+1}\right)\right] \\
& = \frac{I_a(A)}{\pi_a\left(\mathbf{O}_{min}; P^*\right)} M_j \left(E_{P^*}\left[\frac{I_a(A)}{\pi_a\left(\mathbf{O}_{min}; P^*\right)}\, M^{**}|\, M_j, \mathrm{pa}_{\mathcal{G}}\left(M_j\right)\right] - M^{**}\right),
\end{aligned}
$$

which is a non-constant function of M_j. ■

References

Alberto Abadie and Matias D Cattaneo. Econometric methods for program evaluation. *Annual Review of Economics*, 10:465–503, 2018.

Steen A Andersson, David Madigan, Michael D Perlman, et al. A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

Norman Breslow. Design and analysis of case-control studies. *Annual review of public health*, 3(1):29–54, 1982.

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.

David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *Journal of Machine Learning Research*, 2(Feb):445–498, 2002.

William G Cochran. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, pages 295–313, 1968.

BL De Stavola and DR Cox. On the consequences of overstratification. *Biometrika*, 95(4): 992–996, 2008.

Marco Eigenmann, Preetam Nandy, and Marloes H. Maathuis. Structure learning of linear gaussian structural equation models with weak edges. In *UAI'17*, 2017.

Robin J Evans. Graphs for margins of Bayesian Networks. *Scandinavian Journal of Statistics*, 43(3):625–648, 2016.

Robin J Evans and Thomas S Richardson. Markovian acyclic directed mixed graphs for discrete data. *The Annals of Statistics*, 42(4):1452–1482, 2014.

Mitchell H Gail. The effect of pooling across strata in perfectly balanced studies. *Biometrics*, pages 151–162, 1988.

Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in Bayesian networks. *Networks*, 20(5):507–534, 1990.

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

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012.

Takahiro Hayashi and Manabu Kuroki. On estimating causal effects based on supplemental variables. In *AISTATS'14*, pages 312–319, 2014.

Leonard Henckel, Emilija Perković, and Marloes H Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv preprint arXiv:1907.02435*, 2019.

Miguel A Hernan and James M Robins. *Causal inference*. CRC Boca Raton, FL, 2019.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Patrik O. Hoyer, Aapo Hyvärinen, Richard Scheines, Peter Spirtes, Joseph Ramsey, Gustavo Lacerda, and Shohei Shimizu. Causal discovery of linear acyclic models with arbitrary distributions. In *UAI'08*, pages 282–289, 2008.

Arthur B Kahn. Topological sorting of large networks. *Communications of the ACM*, 5 (11):558–562, 1962.

Manabu Kuroki and Zhihong Cai. Selection of identifiability criteria for total effects by using path diagrams. In *UAI'04*, pages 333–340, 2004.

Manabu Kuroki and Masami Miyakawa. Covariate selection for estimating the causal effect of control plans by using causal diagrams. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):209–222, 2003.

Steffen L Lauritzen. *Graphical Models*. Clarendon Press, 1996.

Marloes H. Maathuis and Diego Colombo. A generalized back-door criterion. *The Annals of Statistics*, 43(3):1060–1088, 06 2015.

Nathan Mantel and William Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22(4):719–748, 1959.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *UAI'95*, pages 403–410, 1995.

Markus Neuhaeuser and Heiko Becher. Improved odds ratio estimation by post hoc stratification of case-control data. *Statistics in medicine*, 16(9):993–1004, 1997.

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

Judea Pearl. *Causality: models, reasoning and inference*. Springer, 2000.

Judea Pearl and James M Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In *UAI'95*, pages 444–453, 1995.

Emilija Perković. Identifying causal effects in maximally oriented partially directed acyclic graphs. *arXiv preprint arXiv:1910.02997*, 2019.

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

James Robins. A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

James M Robins. Addendum to a new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Computers & Mathematics with Applications*, 14(9-12):923–945, 1987a.

James M Robins. Addendum to a new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Computers & Mathematics with Applications*, 14(9-12):923–945, 1987b.

James M Robins and Thomas S Richardson. Alternative graphical causal models and the identification of direct effects. *Causality and Psychopathology: Finding the determinants of disorders and their cures*, pages 103–158, 2010.

James M Robins and Andrea Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*, pages 297–331. Springer, 1992.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

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

Laurence D Robinson and Nicholas P Jewell. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review/Revue Internationale de Statistique*, pages 227–240, 1991.

Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(Sep):1941–1979, 2008.

Ilya Shpitser, Tyler VanderWeele, and James M. Robins. On the validity of covariate adjustment for estimating causal effects. In *UAI'10*, pages 527–536, 2010.

Ilya Shpitser, Robin J Evans, Thomas S Richardson, and James M Robins. Introduction to Nested Markov models. *Behaviormetrika*, 41(1):3–39, 2014.

Ezequiel Smucler, Andrea Rotnitzky, and James M Robins. A unifying approach for doubly-robust l_1 regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*, 2019.

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

Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *UAI'02*, pages 519–527, 2002.

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

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

Aad van der Vaart. Higher order tangent spaces and influence functions. *Statistical Science*, pages 679–686, 2014.

Aad W Van der Vaart. *Asymptotic statistics*. Cambridge university press, 2000.

Benito van der Zander and Maciej Liskiewicz. Finding minimal d-separators in linear time and applications. In *UAI'19*, pages 637–647, 2019.

Tyler J VanderWeele and Ilya Shpitser. A new criterion for confounder selection. *Biometrics*, 67(4):1406–1413, 2011.

Thomas Verma and Judea Pearl. Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition*, volume 9, pages 69–76. Elsevier, 1990.

Yuhao Wang, Liam Solus, Karren Dai Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In *NIPS'17*, pages 5824–5833, 2017.

Janine Witte, Leonard Henckel, Marloes H. Maathuis, and Vanessa Didelez. On efficient adjustment in causal graphs. *arXiv e-prints*, 2020.