Journal de la Société Française de Statistique
Vol. 161 No. 1 (2020)

Identification in Causal Models With Hidden Variables

Ilya Shpitser[1]

Abstract: Targets of inference that establish causality are phrased in terms of counterfactual responses to interventions. These *potential outcomes* operationalize cause effect relationships by means of comparisons of cases and controls in hypothetical randomized controlled experiments. In many applied settings, data on such experiments is not directly available, necessitating assumptions linking the counterfactual target of inference with the factual observed data distribution. This link is provided by causal models. Originally defined on potential outcomes directly (Rubin, 1976), causal models have been extended to longitudinal settings (Robins, 1986), and reformulated as graphical models (Spirtes et al., 2001; Pearl, 2009). In settings where common causes of all observed variables are themselves observed, many causal inference targets are identified via variations of the expression referred to in the literature as the *g-formula* (Robins, 1986), the *manipulated distribution* (Spirtes et al., 2001), or the *truncated factorization* (Pearl, 2009).

In settings where hidden variables are present, identification results become considerably more complicated. In this manuscript, we review identification theory in causal models with hidden variables for common targets that arise in causal inference applications, including causal effects, direct, indirect, and path-specific effects, and outcomes of dynamic treatment regimes. We will describe a simple formulation of this theory (Tian and Pearl, 2002; Shpitser and Pearl, 2006b,a; Tian, 2008; Shpitser, 2013) in terms of causal graphical models, and the fixing operator, a statistical analogue of the intervention operation (Richardson et al., 2017).

Keywords: identification, graphical models, causal inference, hidden variable models
AMS 2000 subject classifications: 62H99, 60E05

1. Introduction

In causal inference, the relationship between causes (termed exposures or treatments) and effects (termed outcomes) is quantified via responses of outcomes to different hypothetical assignments of values to causes. These responses are called *potential outcomes*, and were first defined in the context of analysis of experimental data in (Neyman, 1923). Randomized treatment assignment in trials of the kind considered by Neyman implies variation in outcomes in response to changes in treatment assignments, detected by standard statistical inference methods, could be used to draw valid causal inferences.

In datasets where treatment assignment is not randomized, causal inference is not always possible due to spurious associations between treatments and outcomes introduced by their unobserved common causes. Nevertheless, certain assumptions, given by causal models, allow a link to be made between data that is actually observed, and causal parameters involving outcomes that would have occurred under hypothetical treatment assignments. An early example of such a link via the *conditionally ignorable model* and the *stable unit treatment value assumption (SUTVA)*

[1] Department of Computer Science, Johns Hopkins University, 3400 N Charles St, Baltimore, MD 21218
E-mail: ilyas@cs.jhu.edu

is found in (Rubin, 1976). More general models which allow identification of causal parameters in longitudinal observational data via the g-computation algorithm formula (g-formula) were developed in (Robins, 1986). Causal models in general settings were reconceptualized in terms of graphical models by Pearl (2009) and Spirtes et al. (2001), with early versions for the case of linear causal relationships given in (Wright, 1921; Haavelmo, 1943). Graphical models were used to derive algorithms which express causal parameters of interest as functions of observed data, and which generalize g-computation (Tian and Pearl, 2002; Shpitser and Pearl, 2006b,a). These algorithms have also been extended to mediation analysis problems, where effects along causal pathways are of interest (Shpitser, 2013; Shpitser and Tchetgen Tchetgen, 2016), and problems, such as those encountered in precision medicine, where responses to treatments are assigned by a policy (Tian, 2008).

In this paper, we review modern non-parametric identification theory for parameters of interest in causal models represented by directed acyclic graphs (DAGs), possibly with hidden variables. We describe how functionals of the observed data corresponding to causal parameters identified under a fully observed DAG model have a close relationship to a truncated version of the Markov factorization associated with DAGs. Similarly, we describe how not every causal parameter is identified if hidden variables are present, and that causal parameters that are identified under a hidden variable DAG model have a close relationship to a truncated version of the *nested* Markov factorization of the observed marginal distribution, associated with a mixed graph representing that hidden variable DAG (Richardson et al., 2017). We reformulate existing identification theory in terms of this factorization, phrased in terms of mixed graphs, kernels (which generalize conditional densities), and a fixing operation which can be viewed as a statistical analogue of interventions, and which generalizes marginalization, conditioning, and applications of the g-formula .

2. Preliminaries

We will associate random variables with vertices in graphs. We will denote *both* a single vertex and a single corresponding random variable as an uppercase Roman letter, e.g. A. Sets of vertices (and corresponding random variables) will be denoted by upper case vectors, e.g. $\vec{A}$. For a random variable V, we denote the state space of V as $\mathfrak{X}_V$. For example if V is binary, then $\mathfrak{X}_V = \{0,1\}$. We denote elements of a set $\mathfrak{X}_A$ (values of A) by lowercase Roman letters: $a \in \mathfrak{X}_A$. The state space of a set $\vec{V}$ of random variables is simply the Cartesian product of the individual state spaces: $\mathfrak{X}_{\vec{V}} = \times_{V \in \vec{V}} (\mathfrak{X}_V)$. Sets of values corresponding to sets of random variables will be denoted by lowercase vectors, e.g. $\vec{a} \in \mathfrak{X}_{\vec{A}}$. Sometimes we will denote a restriction of a set of values by a set subscript. That is if $\vec{v}$ is a set of values of $\vec{V}$, and $\vec{A} \subseteq \vec{V}$, then $\vec{v}_{\vec{A}}$ is a restriction of $\vec{v}$ to values in $\vec{A}$.

2.1. *The Intervention Operation*

Just like statistical models are sets of joint distributions representing uncertainty about an observable situation, causal models are sets of joint distributions representing uncertainty about a *counterfactual situation*. Counterfactual situations are represented by means of an *intervention operation* (Pearl, 2009), and hypothetical responses to this operation (Neyman, 1923).

For a subset $\vec{A}$ of random variables $\vec{V}$, and a value assignment $\vec{a}$ to $\vec{A}$, an intervention is a forced assignment of $\vec{A}$ to an element of $\mathfrak{X}_{\vec{A}}$. The intervention operation which maps $\vec{A}$ to $\vec{a} \in \mathfrak{X}_{\vec{A}}$ was denoted by $\mathrm{do}(\vec{a})$ by Pearl (2009). The result of the intervention operation $\mathrm{do}(\vec{a})$ is a counterfactual world "minimally altered" from the factual world such that variables in $\vec{A}$ have values $\vec{a}$. If the factual world is represented by a set of factual random variables $\vec{V}$, and their joint distribution $p(\vec{V})$, the counterfactual world is represented by a set of potential outcome random variables $\{V(\vec{a}) \mid V \in \vec{V} \setminus \vec{A}\}$, and their joint distribution written as $p(\vec{V} \setminus \vec{A} \mid \mathrm{do}(\vec{a}))$. In general, the intervention operation $\mathrm{do}(\vec{a})$ does not correspond to conditioning on the event $\vec{A} = \vec{a}$. To define potential outcomes, and the notion of "minimal alteration" formally, we first introduce graphs, graphical models, and causal models.

2.2. *Acyclic Directed Graphs*

We will define causal models using directed graphs. A directed graph only contains directed edges ($\rightarrow$). We will denote graphs by capital calligraphy letters $\mathscr{G}$, and when necessary will explicitly add their vertex sets as part of the notation: $\mathscr{G}(\vec{V})$. A directed graph may contain at most one edge between two vertices.

We denote edges as ordered pairs of distinct vertices subscripted by the type of edge. For example $(AB)_{\rightarrow}$ is a directed edge from A to B. We omit the subscript when the edge type is not relevant. A path is a sequence of edges of the form $\langle (V_1V_2), (V_2V_3), \ldots, (V_{k-2}V_{k-1}), (V_{k-1}V_k) \rangle$, where $V_1 \neq V_k$. Edges in a path may only occur once, and vertices may appear at most twice as elements of edges that are adjacent on the path. We will denote paths by indexed Greek letters, e.g. π_i, and sets of paths by Greek letters, e.g. π.

A directed path from V_1 to V_k has the form $\langle (V_1V_2)_{\rightarrow}, (V_2V_3)_{\rightarrow}, \ldots, (V_{k-2}V_{k-1})_{\rightarrow}, (V_{k-1}V_k)_{\rightarrow} \rangle$. A directed graph $\mathscr{G}$ has a directed cycle if it contains a path of the form $\langle (V_1V_2)_{\rightarrow}, (V_2V_3)_{\rightarrow}, \ldots, (V_{k-2}V_{k-1})_{\rightarrow}, (V_{k-1}V_k)_{\rightarrow} \rangle$, and an edge $(V_k, V_1)_{\rightarrow}$. A directed acyclic graph (DAG) is a directed graph with no directed cycles. Given a DAG $\mathscr{G}(\vec{V})$, and a subset $\vec{A} \subset \vec{V}$, define the induced subgraph $\mathscr{G}(\vec{V})_{\vec{A}}$ as the DAG containing vertices $\vec{A}$ and any edge in $\mathscr{G}(\vec{V})$ between elements in $\vec{A}$.

Given a DAG $\mathscr{G}$, define the following genealogic sets: *parents, children, ancestors, and descendants* of V, to be

$$\begin{aligned}
\mathrm{pa}_{\mathscr{G}}(V) &\equiv \{Z \in \vec{V} \mid (ZV)_{\rightarrow} \text{ exists in } \mathscr{G}\}, \\
\mathrm{ch}_{\mathscr{G}}(V) &\equiv \{Z \in \vec{V} \mid (VZ)_{\rightarrow} \text{ exists in } \mathscr{G}\}, \\
\mathrm{an}_{\mathscr{G}}(V) &\equiv \{Z \in \vec{V} \mid \langle (ZV_1)_{\rightarrow}, \ldots, (V_kV)_{\rightarrow} \rangle \text{ exists in } \mathscr{G}\}, \\
\mathrm{de}_{\mathscr{G}}(V) &\equiv \{Z \in \vec{V} \mid \langle (VV_1)_{\rightarrow}, \ldots, (V_kZ)_{\rightarrow} \rangle \text{ exists in } \mathscr{G}\}.
\end{aligned}$$

Define the set of *non-descendants* of $V \in \vec{V}$ to be $\mathrm{nd}_{\mathscr{G}}(V) \equiv \vec{V} \setminus \mathrm{de}_{\mathscr{G}}(V)$. By convention, for any $V \in \vec{V}$, $V \in \mathrm{an}_{\mathscr{G}}(V) \cap \mathrm{de}_{\mathscr{G}}(V)$. A total ordering $\prec$ on elements in $\vec{V}$ in a DAG $\mathscr{G}$ is called topological with respect to $\mathscr{G}$ if for all distinct $V_1, V_2 \in \vec{V}$, whenever $V_1 \prec V_2$, $V_1 \notin \mathrm{de}_{\mathscr{G}}(V_2)$.

DAGs have been used to define a statistical model via sets of conditional independence statements represented by Markov properties and a factorization. We review this model as a prelude for the more complex definition of a *causal model* of a DAG.

2.3. Statistical Models Of A DAG

A statistical model of a DAG $\mathscr{G}(\vec{V})$, or a Bayesian network, is a set of distributions $p(\vec{V})$ such that

$$p(\vec{V}) = \prod_{V \in \vec{V}} p(V \mid \mathrm{pa}_{\mathscr{G}(\vec{V})}(V)). \tag{1}$$

Any distribution in the above set is said to Markov factorize or be Markov relative to $\mathscr{G}(\vec{V})$. An alternative formulation of statistical models of a DAG is in terms of the global Markov property defined via d-separation (Pearl, 1988), which we reproduce below.

A sequence of two consecutive edges on a path is called a *triple*. A triple of the form $\langle (AC)_{\to} (CB)_{\leftarrow} \rangle$ is called a collider, a triple of the form $\langle (AC)_{\to} (CB)_{\to} \rangle$ is called a chain, and a triple of the form $\langle (AC)_{\leftarrow} (CB)_{\to} \rangle$ is called a fork. The latter two types of triples are collectively called non-colliders. Given a DAG $\mathscr{G}(\vec{V})$, disjoint vertices $A, B \in \vec{V}$, and a set $\vec{C}$, such that $A, B \not\in \vec{C}$, a path from A to B in $\mathscr{G}(\vec{V})$ is said to be *d-separated* given a set $\vec{C}$ if there exists a collider $\langle (AC)_{\to}, (CB)_{\leftarrow} \rangle$ such that $\mathrm{de}_{\mathscr{G}(\vec{V})}(C) \cap \vec{C} = \emptyset$ or a non-collider $\langle (AC), (CB) \rangle$ such that $C \in \vec{C}$. For three disjoint sets $\vec{A}, \vec{B}, \vec{C}$, we say $\vec{A}$ is d-separated from $\vec{B}$ given $\vec{C}$ in $\mathscr{G}(\vec{V})$, written $(\vec{A} \perp\!\!\!\perp \vec{B} \mid \vec{C})_{\mathscr{G}}$ as a shorthand, if every path from any $A \in \vec{A}$ to any $B \in \vec{B}$ is d-separated by $\vec{C}$. The d-separation criterion states d-separation always implies conditional independence in statistical DAG models. That is, if $p(\vec{V})$ is Markov relative to a DAG $\mathscr{G}(\vec{V})$, the following implication holds:

$$\text{if } (\vec{A} \perp\!\!\!\perp \vec{B} \mid \vec{C})_{\mathscr{G}(\vec{V})} \text{ then } (\vec{A} \perp\!\!\!\perp \vec{B} \mid \vec{C})_{p(\vec{V})},$$

where $(\vec{A} \perp\!\!\!\perp \vec{B} \mid \vec{C})_{p(\vec{V})}$ is a shorthand for the statement "$\vec{A}$ is independent of $\vec{B}$ conditional on $\vec{C}$ in $p(\vec{V})$." The global Markov property given by d-separation and the Markov factorization (1) are *equivalent* ways of defining the statistical DAG model, without requiring that elements $p(\vec{V})$ of the model be positive distributions. Such a requirement is necessary for the Hammersley-Clifford theorem for undirected graphical models. See (Lauritzen, 1996) for details.

2.4. Atomic Potential Outcomes and Causal Models Of A DAG

Just as a statistical model of $\mathscr{G}(\vec{V})$ is a set of distributions $p(\vec{V})$ defined by (1), a causal model of $\mathscr{G}(\vec{V})$ is a set of distributions over (atomic) potential outcome random variables in the set

$$\mathbb{V} \equiv \{ V(\vec{v}_{\mathrm{pa}_{\mathscr{G}}(V)}) = f_V(\vec{v}_{\mathrm{pa}_{\mathscr{G}}(V)}, \varepsilon_V) \mid V \in \vec{V}, \text{ any set of values } \vec{v} \text{ of } \vec{V} \} \tag{2}$$

defined by some restrictions[1]. Here, f_V is a structural equation which maps values $\vec{v}_{\mathrm{pa}_{\mathscr{G}}(V)}$ of $\mathrm{pa}_{\mathscr{G}}(V)$ and a random variable ε_V, representing unobserved exogenous factors, to values v of V. Thus, f_V, $\vec{v}$ and the random variable ε_V induce the random variable $V(\vec{v}_{\mathrm{pa}_{\mathscr{G}}(V)})$.

We consider a causal model of $\mathscr{G}$, described in (Pearl, 2009) and (Richardson and Robins, 2013). The *functional model*, also known as the *multiple worlds model* (Shpitser and Tchetgen Tchetgen, 2016), or the *non-parametric structural equation model with independent errors*

[1] To avoid measure-theoretic complications, we consider finite state spaces here.

(NPSEM-IE) is the set of all distributions $p(\mathbb{V})$ such that variables in

$$\left\{ \{ V(\vec{a}_{\text{pa}_{\mathscr{G}}(V)}) : \vec{a}_{\text{pa}_{\mathscr{G}}(V)} \text{ any set of values of } \text{pa}_{\mathscr{G}}(V) \} : V \in \vec{V} \right\} \tag{3}$$

are mutually independent. The assumption (3) corresponding to the functional model can be interpreted to mean that the joint distribution $p(\{\varepsilon_V \mid V \in \vec{V}\})$ factorizes as $\prod_{V \in \vec{V}} p(\varepsilon_V)$. The name "non-parametric structural equation model with independent errors" comes from this property, and the fact the structural equations f_V are unrestricted. As an example, the binary functional model associated with the DAG in Fig. 1 (a) asserts that sets of random variables $\{W\}, \{A(w) \mid w \in \{0,1\}\}, \{M(a,w) \mid a \in \{0,1\}, w \in \{0,1\}\}, \{Y(a,m,w) \mid a \in \{0,1\}, m \in \{0,1\}\}$ are mutually independent.

Weaker models than the functional model exist, such as the *finest fully randomized causally interpreted structured tree graph (FFRCISTG) model*, also known as the *single world model* (Robins, 1986; Shpitser and Tchetgen Tchetgen, 2016; Richardson et al., 2017). In the interests of space, we will not discuss this model further in this paper, although its assumptions generally suffice for identification of any targets described in this paper other than path-specific effects.

2.5. *Defining Arbitrary Potential Outcomes*

Given a set of atomic potential outcomes of the form $V(\vec{v}_{\text{pa}_{\mathscr{G}}(V)})$, other potential outcomes of the form $V(\vec{a})$, where $\vec{A}$ is an arbitrary subset of $\vec{V}$, can be defined by means of *recursive substitution*:

$$Y(\vec{a}) \equiv Y(\vec{a}_{\text{pa}_{\mathscr{G}}(Y) \cap \vec{A}}, \{W(\vec{a}) \mid W \in \text{pa}_{\mathscr{G}}(Y) \setminus \vec{A}\}) \tag{4}$$

In words, this states that the response of Y had we applied the intervention operator $\text{do}(\vec{a})$, that is had we, possibly contrary to fact, set values of $\vec{A}$ to $\vec{a}$, is defined as the potential outcome for Y where

(a) all parents of Y which are in $\vec{A}$ are counterfactually assigned an appropriate value from $\vec{a}$, and

(b) all other parents $W \in \text{pa}_{\mathscr{G}}(Y) \setminus \vec{A}$ are counterfactually assigned whatever value they would have attained had we applied the intervention operator $\text{do}(\vec{a})$.

Responses involving W are themselves counterfactual and defined recursively. The definition terminates because of the lack of directed cycles in $\mathscr{G}$. For example, in the graph in Fig. 1 (a), $Y(a) = Y(a, M(a,W), W)$. The appearance of W as a capital letter in the expression is interpreted to mean "counterfactually assign W to whatever value is actually observed." Recursively setting all ancestors of Y to their actually observed values results in the factual random variable $Y \equiv Y(A, M(A,W), W)$. More generally, every factual variable $V \in \vec{V}$ that is a part of the observed data distribution $p(\vec{V})$ can be formed in this way.

The causal model may impose restrictions on counterfactual variables defined via (4). For example, it is a straightforward consequence of (4) that for any $\vec{A} \subseteq \vec{V}$, such that $\text{pa}_{\mathscr{G}}(V) \subseteq \vec{A}$, and any $\vec{a}_1, \vec{a}_2 \in \mathfrak{X}_{\vec{A} \setminus \text{pa}_{\mathscr{G}}(V)}$, $\vec{a}_0 \in \mathfrak{X}_{\text{pa}_{\mathscr{G}}(V)}$, $V(\vec{a}_0, \vec{a}_1) = V(\vec{a}_0, \vec{a}_2)$. In other words, V is not affected by interventions on any variable outside the set $\text{pa}_{\mathscr{G}}(V)$, as long as all variables in the set $\text{pa}_{\mathscr{G}}(V)$ are already intervened on. In this sense, the variables in $\text{pa}_{\mathscr{G}}(V)$ can be viewed as the *observed direct causes* of V, and the structural equation f_V as the causal mechanism determining the value of V in terms of values of observed direct causes, and the exogenous direct cause ε_V.

The causal model may impose restrictions on factual variables as well. In particular, for the set $\vec{V}$ of factual variables obtained via (4) from atomic counterfactuals $\mathbb{V}$, it is known that if $p(\mathbb{V})$ lies in the functional causal model of $\mathscr{G}(\vec{V})$, then $p(\vec{V})$ obeys (1), in other words $p(\vec{V})$ lies in the statistical model of $\mathscr{G}(V)$. As a matter of convenience, and to avoid complications with identifiability, we will assume positive observed data distributions $p(\vec{V})$ from this point on.

2.6. The Fundamental Problem Of Causal Inference

Most random variables defined via (4) are counterfactual, meaning realizations from these variables are not available as observed data. Observed data is restricted to realizations of the observed distribution $p(\vec{V})$. Making inferences about any counterfactual parameter thus entails building a link between factual and counterfactual variables.

A standard assumption which provides this link is called *consistency*, and states that for any $\vec{a} \in \mathfrak{X}_{\vec{A}}, \vec{b} \in \mathfrak{X}_{\vec{B}}$, and Y, $\vec{A}(\vec{b}) = \vec{a}$ implies $Y(\vec{b}) = Y(\vec{a},\vec{b})$. Though consistency (in the simple form where $\vec{b}$ is empty) is often stated as a standalone part of the *stable unit treatment value assumption (SUTVA)* (Rubin, 1976), it is also a logical consequence of (4) above, see (Malinsky et al., 2019) for a simple proof. In words, consistency states that in any given counterfactual situation where $\vec{B}$ were intervened on values $\vec{b}$, if it were the case that random variables $\vec{A}(\vec{b})$ were observed to attain values $\vec{a}$, then any counterfactual random variable $Y(\vec{b})$ in that situation is equal to the counterfactual random variable $Y(\vec{a},\vec{b})$ representing the situation where $\vec{A}$ were also intervened on to the values $\vec{a}$. Viewed in terms of structural equations, consistency asserts a kind of *mechanism modularity*: for every V, f_V reliably maps inputs to outputs regardless of the type of process that may have set those inputs.

Given an observed dataset specifying values y_i, a_i, for $i = 1,\dots,n$, consistency allows inferences to be made about $Y(a_i)$ from realizations $Y_j(a_i)$ where j is such that $a_j = a_i$. However, for all rows j, realizations of $Y(a_i)$ are missing from the dataset for row j, if $a_i \neq a_j$. This means that for any row j, only half of the potential outcomes are available (possibly less if A has more than two values). This missing data problem is referred to as the *fundamental problem of causal inference* (Rubin, 1976). Making inferences using counterfactual realizations missing in observed data is not always possible, and when it is, relies on additional assumptions over and above consistency, given by a causal model.

2.7. Identification

If assumptions imposed by a particular causal model imply a counterfactual quantity is a unique functional of the observed data distribution in every element of the model, we say the counterfactual is identified in the model. Here we concentrate on identification of counterfactual distributions. Formally, a distribution $\tilde{p}$ or parameter $\tilde{\theta}$ is said to be identified from $p(\vec{V})$ in a causal model, if there exists a function of $p(\vec{V})$ which yields $\tilde{p}$ (or $\tilde{\theta}$) in every element of the model. Identification in this sense is important to establish before estimating $\tilde{p}$ from observed data. A distribution $\tilde{p}$ or parameter $\tilde{\theta}$ is said to be non-identified from $p(\vec{V})$ in a causal model if there exist two elements in the model which share $p(\vec{V})$ but differ in $\tilde{p}$ (or $\tilde{\theta}$). Estimation for non-identified parameters or distributions is an ill-posed problem.

3. Targets of Inference

We now describe common targets of inference, phrased in terms of potential outcome random variables, that arise in causal inference applications.

3.1. Average Treatment Effects

The relationship of a set of treatments $\vec{A}$ and an outcome Y is quantified via a comparison of potential outcomes $Y(\vec{a})$, $Y(\vec{a}')$ under two hypothetical value assignments $\vec{a}$ and $\vec{a}'$ to $\vec{A}$, representing cases and controls. These comparisons are often made on the mean difference scale. This contrast is known as the *average causal effect* (ACE):

$$\mathbb{E}[Y(\vec{a})] - \mathbb{E}[Y(\vec{a}')].$$

For a set of treatments $\vec{A}$, a comparison of interest might include the *controlled direct effect*, where outcomes Y are compared for two treatment trajectories $\vec{a} \in \mathfrak{X}_{\{A_1,\ldots,A_k\}}$ that agree on all values $\vec{a}_{-i} \equiv \{a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_k$ except values a_i, a_i':

$$\mathbb{E}[Y(\vec{a}_{-i},a_i)] - \mathbb{E}[Y(\vec{a}_{-i},a_i')].$$

Identifiable average causal effects and controlled direct effects, possibly also conditioned on a vector of baseline covariates, are sometimes modeled directly in semi-parametric approaches based on structural nested models and marginal structural models (Robins, 1999a,b).

3.2. Mediation Analysis And Path-Specific Effects

Given that a treatment effect of A on Y is established, it is often desirable to understand the mechanism by which the causal influence takes place. A simple type of mechanisms analysis is mediation analysis – the decomposition of the treatment effect into components associated with causal pathways from A to Y.

While controlled direct effects capture a type of direct effect of A on Y (that is, the effect not mediated by other variables), it suffers from two disadvantages. First, since other direct causes are fixed to a set of reference values, the controlled direct effect is best viewed not as a single contrast, but a potentially high dimensional mapping from the values of other direct causes of the outcome to contrasts. This makes this type of effect potentially difficult to estimate and interpret. Second, there is no corresponding version of the *indirect effect*.

An alternative, proposed in (Robins and Greenland, 1992; Pearl, 2001), is to define *natural* direct and indirect effects by means of nested counterfactuals of the form $Y(a,M(a'))$, and $Y(a',M(a))$, where a is the treatment value corresponding to cases, and a' the treatment value corresponding to controls. These direct and indirect effects are defined in such a way that they form a decomposition of the average treatment effect on the appropriate scale. For example, the following definitions give an additive decomposition of the average treatment effect defined on the mean difference scale:

$$\begin{aligned}\mathbb{E}[Y(a)] - \mathbb{E}[Y(a')] &= \underbrace{(\mathbb{E}[Y(a)] - \mathbb{E}[Y(a,M(a'))])}_{\text{total indirect effect}} + \underbrace{(\mathbb{E}[Y(a,M(a'))] - \mathbb{E}[Y(a')])}_{\text{pure direct effect}} \\ &= \underbrace{(\mathbb{E}[Y(a)] - \mathbb{E}[Y(a',M(a))])}_{\text{total direct effect}} + \underbrace{(\mathbb{E}[Y(a',M(a))] - \mathbb{E}[Y(a')])}_{\text{pure indirect effect}}.\end{aligned}$$

One way of conceptualizing these types of effects (Robins and Richardson, 2010) is to assume a treatment A can be partitioned into two components: a component that only directly influences the outcome Y, and a component that only directly influences a mediator variable M. In the toy example discussed in (Robins and Richardson, 2010), we may believe smoking affects health due to the presence of smoke in the lungs, which may cause cancer, and by means of nicotine in the bloodstream, which may contribute to heart disease. If we are interested in the effect of smoking mediated by cancer, we may consider comparing smokers to a population which only received the component of the treatment where the influence of smoke is absent, and the influence of nicotine is present, such as a nicotine patch. If we view Y as health status, M as cancer, and A as smoking, then the expected value contrast above corresponding to the total indirect effect of smoking would represent the comparison of health outcomes in smokers and nicotine patch users.

The intuition behind direct and indirect effects can be generalized to settings where an effect along a particular causal path is of interest. Consider an example from (Miles et al., 2017) represented by Fig. 1 (b), representing a cross-sectional study of HIV patients. Here, W is a set of baseline characteristics, A is exposure to one of two HIV treatments, M is treatment toxicity, Z is a measure of treatment adherence (how much of the prescribed treatment the patient actually took), and Y is the outcome such as viral failure. In this example, we may be interested in assessing not only the effectiveness of the drug itself, as quantified by the average causal effect, but also the extent to which the drug might influence the outcome through adherence but not toxicity. This type of adherence might be affected by the size of the pill, if the drug is ingested in pill form, or a patient in a low food security situation being prescribed a drug that must be taken with a meal. The influence of the drug on the outcome involved in this type of adherence corresponds to the influence of A on Y along the path $\langle (AZ)_{\rightarrow}, (ZY)_{\rightarrow} \rangle$. Another issue of possible interest is lack adherence due to toxicity of the drug. The influence of A on Y involved in this type of adherence corresponds to the influence of A on Y along the path $\langle (AM)_{\rightarrow}, (MZ)_{\rightarrow}, (ZY)_{\rightarrow} \rangle$.

Effects of treatments on outcomes along a predefined set of paths are called path-specific effects (Pearl, 2001). Such effects are defined using a special type of potential outcome, where the treatment is set to one value a with respect to variables on the set of paths of interest, and to another value a' with respect to variables on all other paths. A single variable may occur in both types paths: those of interest, and those that are not of interest. We can modify (4) to define these types of potential outcomes as follows. Fix a set of directed paths π from A to Y, and let $\text{pa}^{\pi}_{\mathscr{G}}(Y)$ be the set of parents of Y along an edge which is a part of a path in π, and $\text{pa}^{\overline{\pi}}_{\mathscr{G}}(Y)$ be the set of all other parents of Y. Given π and values a, a', define the π-specific potential outcome Y as

$$\begin{aligned}&Y(\pi,a,a') \equiv a \text{ if Y = A} \\ &Y(\pi,a,a') \equiv Y(\{W(\pi,a,a') \mid W \in \text{pa}^{\pi}_{\mathscr{G}}(Y)\}, \{W(a') \mid W \in \text{pa}^{\overline{\pi}}_{\mathscr{G}}(Y)\})\end{aligned} \tag{5}$$

where $W(a') \equiv a'$ if $W = A$. This definition says that the π-specific potential outcome $Y(\pi, a, a')$ is defined by setting values of $\text{pa}_{\mathscr{G}}(Y)$ using the following four cases:

- If $A \in \text{pa}_{\mathscr{G}}(Y)$ and $(AY)_{\rightarrow}$ is in a path in π, set A to a.
- If $A \in \text{pa}_{\mathscr{G}}(Y)$ and $(AY)_{\rightarrow}$ is not in a path in π, set A to a'.
- If $W \in \text{pa}_{\mathscr{G}}(Y) \setminus \{A\}$ and $(WY)_{\rightarrow}$ is not in any path in π, set W to the value it would have attained under potential outcomes $W(a')$.
- If $W \in \text{pa}_{\mathscr{G}}(Y) \setminus \{A\}$ and $(WY)_{\rightarrow}$ is in some path in π, set W to the value it would have attained under the (recursively defined) π-specific potential outcome $W(\pi, a, a')$.

As was the case with (4), the inductive definition terminates because $\mathscr{G}$ is acyclic. This definition is a natural generalization of the way natural direct effects were defined above, with the direct path $\langle (AY)_{\rightarrow} \rangle$ being a path of interest, and the indirect path $\langle (AM)_{\rightarrow}, (MY)_{\rightarrow} \rangle$ being a path not of interest.

Applying this definition to π consisting of a single path $\langle (AZ)_{\rightarrow}, (ZY)_{\rightarrow} \rangle$ in Fig. 1 (b) yields

$$Y(\pi, a, a') \equiv Y(a', Z(a, M(a', W)), M(a', W), W).$$

By analogy with direct and indirect effects, we can use this counterfactual to define the total effect not through π as

$$\mathbb{E}[Y(a)] - \mathbb{E}[Y(\pi, a, a')] = \mathbb{E}[Y(a)] - \mathbb{E}[Y(a', Z(a, M(a', W)), M(a', W), W)].$$

and the pure π-specific effect as

$$\mathbb{E}[Y(\pi, a, a')] - \mathbb{E}[Y(a')] = \mathbb{E}[Y(a', Z(a, M(a', W)), M(a', W), W)] - \mathbb{E}[Y(a')].$$

Equation (5) generalizes in a natural way to multiple treatments $\vec{A}$, and multiple outcomes $\vec{Y}$. In such cases, attention is restricted to sets π of *proper causal paths* for $\vec{A}$ and $\vec{Y}$, which are directed paths from an element in $\vec{A}$ to any element in $\vec{Y}$ that otherwise does not intersect $\vec{A}$. In such cases, definition (5) is applied to every $Y \in \vec{Y}$, with a set of proper causal paths π, a set of case values $\vec{a}$, and a set of control values $\vec{a}'$. The resulting distribution is $p(\{Y(\pi, \vec{a}, \vec{a}') \mid Y \in \vec{Y}\})$.

3.3. Dynamic Treatment Regimes

All targets considered so far represented potential outcomes where treatments were set to specific constant values, representing case and control treatments. Distributions over these outcomes can be used to make quantitative comparisons of how different treatments affect average responses in a population. However, in settings such as precision medicine, the primary goal is obtaining good outcomes for every individual, rather than assessing average treatment effects in a population. This difference in crucial where treatments can cause allergies and side effects – a treatment may be both very effective on average and extremely harmful for certain subsets of patients.

In simple versions of this setting, shown in Fig. 1 (a), the goal is not to learn the potential outcome $Y(a)$ given a particular treatment assignment a, but the potential outcome $Y(A = g(W))$, where A is counterfactually set not to a fixed value a, but to a value given by a policy or *dynamic treatment regime* $g(W)$ that depends on the vector of baseline factors W (Chakraborty and Moodie, 2013). Policy quality can be assessed by comparing two different policies, as was done with treatment effects, or finding the optimal policy directly. Just as was the case with $Y(a)$, the response $Y(A = g(W))$ to a dynamic treatment regime is a counterfactual random variable, and is not necessarily identified from observed data, as we will see below. This is because assignment to A given W in the observed data is not necessarily according to the policy of interest $g(.)$. There is a close relationship between the distinction between factual and counterfactual policies and *off-policy learning* in the reinforcement learning literature.

More generally, given a temporally ordered set of treatments $\vec{A} = \{A_1, \ldots, A_k\}$, assessing the quality of a set of policies $\vec{g}_{\vec{A}} = \{g_{A_1}, \ldots, g_{A_k}\}$ with respect to a response Y might be of interest. In such settings, complex interdependence between policies is possible. For a simple example, consider Fig. 1 (b), where W is a set of baseline factors, A, Z are treatments of interest, M is an intermediate outcome, and Y is the final outcome. We assume a temporal order W, A, M, Z, Y on the variables, and allow the policy determining the value of each treatment to depend on the entire observed history, that is g_A is a function of W and g_Z is a function of M, A, W. In medical settings, A may represent first line treatments, and Z secondary treatment options given poor response to the first line treatment. The interdependence of policies g_A and g_Z occurs because g_Z depends on the entire observed history up to Z, and specifically on M, which itself depends on g_A. Defining $Y(\vec{g}_{\vec{A}})$ entails recursively setting all parents of Y to values they would have attained under $\vec{g}_{\vec{A}}$, with the result being the following generalization of (4):

$$\begin{aligned} Y(\vec{g}_{\vec{A}}) &\equiv Y(A = g_A(W), Z = g_Z(A = g_A(W), M = M(\vec{g}_{\vec{A}}), W), W) \\ M(\vec{g}_{\vec{A}}) &\equiv M(A = g_A(W), W), \end{aligned} \tag{6}$$

In oncology, an example corresponding to Fig. 1 (b) would represent assigning treatments to cancer patients, with A representing types of induction chemotherapy, and Z being either continuation of induction chemotherapy, or a switch to salvage chemotherapy in patients that are not responding to primary induction chemotherapy. Naturally, a policy g_Z which switches treatments effectively would depend on the intermediate outcome $M(A = g_A(W), W)$, which measures the degree of response to induction chemotherapy. This intermediate outcome would itself depend on the choice of induction therapy. This choice, in turn, may be governed by patient age, and other baseline covariates that form a part of W.

The definition of the response Y to an arbitrary set of policies is as follows. Given a set of treatments $\vec{A}$ in a causal model represented by a DAG $\mathscr{G}(\vec{V})$, fix a topological ordering $\prec$ on $\vec{V}$ consistent with $\mathscr{G}$ (typically the temporal order for variables in the data), and consider for every $A \in \vec{A}$ a set of variables $\vec{W}_A$ earlier in the ordering $\prec$ than A. Fix a set of policies $\vec{g}_{\vec{A}} \equiv \{g_A(\vec{W}_A) \mid A \in \vec{A}\}$ that determine the value of each $A \in \vec{A}$ using values of $\vec{W}_A$. For any $Y \in \vec{V} \setminus \vec{A}$, the counterfactual response Y had every $A \in \vec{A}$ been determined by $\vec{g}_{\vec{A}}$ is defined using the appropriate generalization of the recursive substitution definition (4):

$$Y(\vec{g}_{\vec{A}}) = Y(\{A = g_A(\vec{W}_A(\vec{g}_{\vec{A}})) | A \in \text{pa}_{\mathscr{G}}(Y) \cap \vec{A}\}, \{W(\vec{g}_{\vec{A}}) | W \in \text{pa}_{\mathscr{G}}(Y) \setminus \vec{A}\}). \tag{7}$$

In words, this says that to define the response of Y given a set of policies $\vec{g}_{\vec{A}}$, we counterfactually set each element of $\text{pa}_{\mathscr{G}}(Y)$ as follows. Each element A of $\text{pa}_{\mathscr{G}}(Y) \cap \vec{A}$ is set to the value of the appropriate g_A evaluated given the values of its inputs $\vec{W}_A$ which are themselves evaluated recursively given $\vec{g}_{\vec{A}}$. Each element W of $\text{pa}_{\mathscr{G}}(Y) \setminus \vec{A}$ is set to a value it would have attained had it been evaluated, recursively, under the policy set $\vec{g}_{\vec{A}}$.

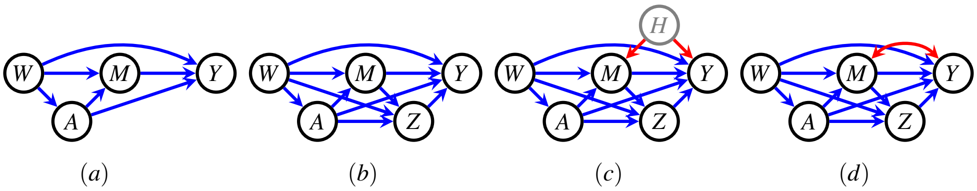

FIGURE 1. *(a) A simple causal DAG, with a single treatment A, a single outcome Y, a vector W of baseline variables, and a single mediator M. (b) A more complex causal DAG with two mediators M and Z. (c) A version of the DAG in (b) with an unobserved confounder H. (d) The ADMG obtained from the DAG in (c) via the latent projection operation collapsing over the unobserved variable H.*

The quality of the policy set $\vec{g}_{\vec{A}}$ is often expressed as the outcome expectation under the policy set: $\mathbb{E}[Y(\vec{g}_{\vec{A}})]$, or any other appropriate function of $Y(\vec{g}_{\vec{A}})$.

4. Identification In Fully Observed Causal Models Of A DAG

Having given the necessary preliminaries, and defined appropriate targets of inference, we now consider the question of their identification. In causal models of a DAG where all variables are observed, interventional distributions of the form $p(\vec{Y} \mid \text{do}(\vec{a}))$ and responses to dynamic treatment regimes are always identified, while path-specific effects are identified according to a simple criterion on the graph, given further below.

4.1. Identification Of Interventional Distributions And Responses To Dynamic Treatment Regimes

For any value set $\vec{a}$ of $\vec{A} \subseteq \vec{V}$, the interventional distribution $p(\vec{V} \setminus \vec{A} \mid \text{do}(\vec{a}))$ over counterfactuals $\{V(\vec{a}) \mid V \in \vec{V} \setminus \vec{A}\}$ is identified by

$$p(\vec{V} \setminus \vec{A} \mid \text{do}(\vec{a})) = \prod_{V \in \vec{V} \setminus \vec{A}} p(V \mid \text{pa}_{\mathscr{G}}(V))|_{\vec{A}=\vec{a}} \tag{8}$$

This equation is known as the *g-formula* (Robins, 1986), the *manipulated distribution* (Spirtes et al., 2001), or the *truncated factorization* (Pearl, 2009).

The g-formula asserts that in the functional model corresponding to a DAG $\mathscr{G}(\vec{V})$, the effect of setting any set of variables $\vec{A}$ to values $\vec{a}$ using the intervention operation $\text{do}(\vec{a})$ amounts to replacing the set of structural equations $\{f_V : \mathfrak{X}_{\text{pa}_{\mathscr{G}}(V) \cup \{\varepsilon_V\}} \to \mathfrak{X}_V \mid V \in \vec{A} \text{ or } \vec{A} \cap \text{pa}_{\mathscr{G}}(V) \neq \emptyset\}$ by another set

$$\{\tilde{f}_V : \mathfrak{X}_{(\text{pa}_{\mathscr{G}}(V) \setminus \vec{A}) \cup \{\varepsilon_V\}} \to \mathfrak{X}_V \mid V \in \vec{V} \setminus \vec{A}\} \cup \{\tilde{f}_A : \emptyset \to \{a\} \mid A \in \vec{A}\},$$

where for every V, and $\vec{w} \in \mathfrak{X}_{\text{pa}_{\mathscr{G}}(V) \setminus \vec{A}}$, $\tilde{f}_V(\vec{w}, \varepsilon_V) = f_V(\vec{w}, \tilde{\vec{a}}, \varepsilon_V)$, where $\tilde{\vec{a}}$ is the subset of values of $\vec{a}$ corresponding to $\text{pa}_{\mathscr{G}}(V) \cap \vec{A}$. In words, $\text{do}(\vec{a})$ is implemented by replacing all structural equations that determine elements $A \in \vec{A}$, by new structural equations $\tilde{f}_A$ that ignore all inputs and set A to the appropriate value for A in $\vec{a}$, and replacing all structural equations f_V that determine elements $V \in \vec{V} \setminus \vec{A}$ by structural equations $\tilde{f}_V$ that agree with f_V, except they always evaluate $\text{pa}_{\mathscr{G}}(V) \cap \vec{A}$ using the appropriate subset of values in $\vec{a}$.

Implementing interventions by replacing structural equations generalizes in a straightforward way to yield responses to a set of dynamic treatment regimes $\vec{g}_{\vec{A}}$. Such responses are generated by replacing $f_A : \mathfrak{X}_{\text{pa}_{\mathscr{G}}(A)} \to \mathfrak{X}_A$, for each $A \in \vec{A}$, not by $\tilde{f}_A : \emptyset \to \{a\}$, but by $g_A : \mathfrak{X}_{\vec{W}_A} \to \mathfrak{X}_A$. Similarly, for each V such that $\text{pa}_{\mathscr{G}}(V) \cap \vec{A} \neq \emptyset$, $\tilde{f}_V : \mathfrak{X}_{(\text{pa}_{\mathscr{G}}(V)\setminus\vec{A})}$ agree with f_V, except they always evaluate $\text{pa}_{\mathscr{G}}(V) \cap \vec{A}$ using the appropriate subset of values of $\vec{A}$ determined by $\vec{g}_{\vec{A}}$. This immediately yields the following expression as the identifying formula for the distribution $p(\{V(\vec{g}_{\vec{A}}) | V \in \vec{V} \setminus \vec{A}\})$ over responses in $\vec{V} \setminus \vec{A}$ to a treatment regime $\vec{g}_{\vec{A}}$

$$\prod_{V \in \vec{V} \setminus \vec{A}} p(V | \text{pa}_{\mathscr{G}}(V) \setminus \vec{A}, \{A = g_A(\vec{W}_A) | A \in \text{pa}_{\mathscr{G}}(V) \cap \vec{A}\}) \tag{9}$$

This formula is a generalization of (8).

The expression (8) has a number of well-known special cases. For instance, in Fig. 1 (a),

$$\begin{aligned} p(Y \mid \text{do}(a)) = \sum_{M,W} p(Y,M,W \mid \text{do}(a)) &= \sum_{M,W} p(Y \mid M,a,W)p(M \mid a,W)p(W) \\ = \sum_{W} p(Y \mid a,W)p(W), \end{aligned} \tag{10}$$

where the last equality follows by marginalizing out M and chain rule. This recovers the well-known adjustment or backdoor formula (Pearl, 2009). In Fig. 1 (b),

$$p(Y \mid \text{do}(a,z)) = \sum_{M,W} p(Y,M,W \mid \text{do}(a,z)) = \sum_{M,W} p(Y \mid a,z,M,W)p(M \mid a,W)p(W).$$

Let g_A be a mapping from values of W to values of A, and g_Z be a mapping from values of M,A,W to values of Z. Then the distribution of the potential outcome $Y(\vec{g}_{\vec{A}}) = Y(\{g_A, g_Z\})$ is identified as

$$p(Y(\vec{g}_{\vec{A}})) = \sum_{W,M} p(Y|A = g_A(W), Z = g_Z(M, A = g_A(W), W), M, W)p(M|A = g_A(W), W)p(W).$$

The first expression recovers the g-computation algorithm formula (Robins, 1986) for inferring causal effects in longitudinal studies, while the second gives the appropriate generalization for dynamic treatment regimes. Both expressions are given here for two time points, but generalize in a straightforward way.

4.2. *Identification Of Path-Specific Effects*

Potential outcomes associated with path-specific effects and defined via (5) are more complicated objects than potential outcomes defined via (4) due to the presence of two conflicting treatment assignments within the same object. A consequence of this is that even in causal models of a DAG $\mathscr{G}(\vec{V})$ where every element of $\vec{V}$ is observed, distributions over some such potential outcomes are not identified. A characterization of identifiable π-specific potential outcomes exists, based on a feature of the graph $\mathscr{G}$ and the path set π (Avin et al., 2005).

Fix a set of treatments $\vec{A}$, a set of outcomes $\vec{Y}$, and a set π of proper causal paths for $\vec{A}$ and $\vec{Y}$ in a DAG $\mathscr{G}(\vec{V})$. A variable $W \in \vec{V} \setminus \vec{A}$ is called a *recanting witness* for π if there exists a directed path in π of the form $\langle (AW)_{\rightarrow}, \ldots, (Z_1Y_1)_{\rightarrow} \rangle$ for some $Y_1 \in \vec{Y}$, and another proper causal path of the form $\langle (AW)_{\rightarrow}, \ldots, (Z_2Y_2)_{\rightarrow} \rangle$ for some $Y_2 \in \vec{Y}$ that is not in π. As an example, in Fig. 1 (b), if we are interested in the path-specific effect of A on Y along a single path $\langle (AM)_{\rightarrow}, (MY)_{\rightarrow} \rangle$, that is if $\pi = \{\langle (AM)_{\rightarrow}, (MY)_{\rightarrow} \rangle\}$, then M is a recanting witness for π, since the path $\langle (AM)_{\rightarrow}, (MZ)_{\rightarrow}, (ZY)_{\rightarrow} \rangle$ is a proper causal path for A and Y, is not an element of π, and has as its first edge $(AM)_{\rightarrow}$ which is also the first edge of $\langle (AM)_{\rightarrow}, (MY)_{\rightarrow} \rangle$. Despite the existence of a recanting witness, the potential outcome $Y(\pi, a, a')$ is still definable via (5), and is equal to $Y(a', Z(a', M(a', W)), M(a, W), W)$. Both potential outcomes $M(a', W)$ and $M(a, W)$ appear in this expression, and this is what prevents identification. This issue arises because the same child M of the intervened on variable A, appears both in a path of interest $\langle (AM)_{\rightarrow}, (MY)_{\rightarrow} \rangle$ and in another path $\langle (AM)_{\rightarrow}, (MZ)_{\rightarrow}, (ZY)_{\rightarrow} \rangle$ that is not of interest, but that was involved in defining the potential outcome.

In fact, identification of potential outcomes $\{Y(\pi, \vec{a}, \vec{a}') \mid Y \in \vec{Y}\}$ involved in a path-specific effect is characterized in DAGs by the absence of a recanting witness. Specifically, given disjoint vertex sets $\vec{A}, \vec{Y}$ in a DAG $\mathscr{G}(\vec{V})$, and a set π of proper causal paths for $\vec{A}$ and $\vec{Y}$, the distribution $p(\{Y(\pi, \vec{a}, \vec{a}') \mid Y \in \vec{Y}\})$ is identified if and only if there are *no* recanting witnesses for π. If the recanting witness does not exist, then the joint counterfactual distribution over variables $\{Y(\pi, \vec{a}, \vec{a}') \mid Y \in \vec{Y}\}$ is identified via a generalization of equation (8) called the *edge g-formula* (Shpitser and Tchetgen Tchetgen, 2016), with an early version appearing in (Avin et al., 2005):

$$p(\{Y(\pi, \vec{a}, \vec{a}') | Y \in \vec{Y}\}) = \sum_{\vec{V} \setminus (\vec{A} \cup \vec{Y})} \prod_{V \in \vec{V} \setminus \vec{A}} p(V | \vec{a}_{\text{pa}^{\pi}_{\mathscr{G}}(V) \cap \vec{A}}, \vec{a}'_{\text{pa}^{\overline{\pi}}_{\mathscr{G}}(V) \cap \vec{A}}, \text{pa}_{\mathscr{G}}(V) \setminus \vec{A}). \tag{11}$$

Just as the ordinary g-formula, the edge g-formula can be viewed as a truncated DAG factorization. However, in the edge g-formula a variable $A \in \vec{A} \cap \text{pa}_{\mathscr{G}}(V)$ in a Markov factor $p(V \mid \text{pa}_{\mathscr{G}}(V))$ can be set to either its value in $\vec{a}$ or its value in $\vec{a}'$, depending on whether the edge $(AV)_{\rightarrow}$ is a part of a path in π or not.

As an example, a recanting witness does not exist for the path-specific effect of A on Y along the set of paths $\pi \equiv \{\langle (AM)_{\rightarrow}, (MY)_{\rightarrow} \rangle; \langle (AM)_{\rightarrow}, (MZ)_{\rightarrow}, (ZY)_{\rightarrow} \rangle\}$ in Fig. 1 (b). The counterfactual distribution $p(Y(\pi, a, a'))$ corresponding to this set of paths is then identified as

$$p(Y(\pi, a, a')) = \sum_{W,M,Z} p(Y \mid M, W, Z, a') p(Z \mid M, W, a') p(M \mid W, a) p(W).$$

In the classical mediation analysis setting shown in Fig. 1 (a), where we are interested in the direct effect of A on Y, in other words in the path set $\pi \equiv \{\langle (AY)_{\rightarrow} \rangle\}$, the counterfactual distribution $p(Y(\{\langle (AY)_{\rightarrow}\}, a, a')) = p(Y(a, M(a')))$ is identified by the edge g-formula as

$$p(Y(a, M(a'))) = \sum_{W,M} p(Y \mid M, W, a) p(M \mid a', W) p(W).$$

If we are interested in the pure direct effect $\mathbb{E}[Y(a, M(a'))] - \mathbb{E}[Y(a')]$, the formula above recovers the well-known *mediation formula* (Pearl, 2011):

$$\mathbb{E}[Y(a, M(a'))] - \mathbb{E}[Y(a')] = \sum_{W,M} \{\mathbb{E}[Y \mid M, W, a] - \mathbb{E}[Y \mid M, W, a']\} p(M \mid a', W) p(W).$$

5. Identification In Causal Models Of A DAG With Hidden Variables

Identification results described so far assumed a causal model of a DAG $\mathscr{G}(\vec{V})$ where every element in $\vec{V}$ corresponds to an observed variable. Unfortunately, this assumption is unrealistic in practice. For example, in observational studies in healthcare, compliance and health status of enrolled patients are generally not completely observed, except through imperfect proxies. This motivates the study of causal models of a DAG $\mathscr{G}(\vec{V} \cup \vec{H})$ where $\vec{V}$ corresponds to observed variables, and $\vec{H}$ to hidden variables.

The presence of hidden variables considerably complicates identification theory. We will describe a simple characterization of identifiable targets of causal inference in hidden variable causal DAGs, based on mixed graphs, and the fixing operation that can be viewed as a statistical version of the intervention operation. This characterization was described in (Richardson et al., 2017) in the context of treatment effects, and generalized for mediation analysis and dynamic treatment regime problems in (Shpitser and Sherman, 2018).

In a fully observed DAG $\mathscr{G}(\vec{V})$, all identified functionals are based on the g-formula (4), which is a truncated version of the Markov factorization of $p(\vec{V})$ associated with $\mathscr{G}(\vec{V})$. In a hidden variable DAG $\mathscr{G}(\vec{V} \cup \vec{H})$, all identified functionals are also based on a truncated version of a Markov factorization. However, this *nested* Markov factorization is of a *marginal* distribution $p(\vec{V})$, and with respect not to the DAG $\mathscr{G}(\vec{V} \cup \vec{H})$, but a special mixed graph we denote $\mathscr{G}(\vec{V})$ obtained by a *latent projection* operation (Verma and Pearl, 1990) from $\mathscr{G}(\vec{V} \cup \vec{H})$. This factorization is defined in terms of objects called kernels that generalize conditional distributions and which can be "put together" to construct the observed marginal $p(\vec{V})$, as well as certain other distributions that correspond to causal targets of inference.

We now give the roadmap for the definitions we will need that will define the nested factorization. First, we describe special mixed graphs called acyclic directed mixed graphs (ADMGs), and their conditional versions (CADMGs), and generalize existing definitions, and genealogic relations defined for DAGs in Section 2.2 to ADMGs and CADMGs. Then we define a special ADMG called a latent projection (Verma and Pearl, 1990) which represents identification theory for an infinite class of "structurally similar" hidden variable DAGs. We then describe how targets of inference can be defined on this ADMG directly in a way that represents the target in any hidden variable DAG in its class. We then describe kernels, which are generalizations of conditional distributions that will represent terms of the nested factorization. Next, we define the fixing operator (Richardson et al., 2017) on graphs and kernels which will be used iteratively to give the nested factorization. The nested factorization will link CADMG and kernel pairs, with the pair derived from $\mathscr{G}(\vec{V})$ and $p(\vec{V})$ via the fixing operator.

Finally, we define the nested factorization of a distribution with respect to an ADMG, reformulate the ID algorithm (Tian and Pearl, 2002; Shpitser and Pearl, 2006b) as a truncated version of this factorization (Richardson et al., 2017), and generalize this formulation to also give identifying functionals for responses to dynamic treatment regimes and path-specific potential outcomes. We also describe the *potential outcomes calculus* (Malinsky et al., 2019), a generalization of Pearl's do-calculus defined in terms of single world intervention graphs (SWIGs) (Richardson and Robins, 2013) and potential outcomes, and show how this calculus may be used to give a complete identification theory for conditional interventional distributions.

5.1. Mixed Graphs

A mixed graph only contains directed ($\to$) and bidirected ($\leftrightarrow$) edges, although other types of mixed graphs have been considered in the literature (Lauritzen, 1996; Drton, 2009; Shpitser, 2015). In mixed graphs representing causal systems, directed edges represent the possible presence of a direct causal relationship between the vertices sharing the edge, while bidirected edges represent the possible presence of a particular type of spurious association between the vertices sharing the edge, as would occur had there been some unnamed and unobserved common cause. A mixed graph may contain at most two edges between two vertices. Moreover, if two vertices have two edges in common, one of them must be directed and one must be bidirected. An acyclic directed mixed graph (ADMG) is a mixed graph with no directed cycles.

A *conditional ADMG (CADMG)* $\mathscr{G}(\vec{V},\vec{W})$ (Richardson et al., 2017) is a graph with a vertex set $\vec{V} \cup \vec{W}$ where for every $W \in \vec{W}$ it is the case that neither $(ZW)_\to$ nor $(ZW)_\leftrightarrow$ edges exist in $\mathscr{G}$ for any $Z \in \vec{V} \cup \vec{W}$. That is, in any CADMG $\mathscr{G}(\vec{V},\vec{W})$, there is no edge of any kind with an arrowhead pointing into any element $W \in \vec{W}$. Although every element in $\vec{W}$ has this property, there may also exist elements $V \in \vec{V}$ such that neither $(ZV)_\to$ nor $(ZV)_\leftrightarrow$ edges exist in $\mathscr{G}(\vec{V},\vec{W})$ for any $Z \in \vec{V} \cup \vec{W}$. The distinction between elements in $\vec{W}$ and elements in $\vec{V}$ does not come from the way these vertex sets are defined, but from how they are used when defining graphical models. Vertices in $\vec{V}$ correspond to random variables, as in the statistical model of a DAG. Vertices in $\vec{W}$ correspond to variables that were fixed to a value. CADMGs will correspond to sets of distributions that depend on values in $\mathfrak{X}_{\vec{W}}$. These distributions, called *kernels*, along with CADMGs, will be used later to concisely express identification in terms of a truncated factorization.

A bidirected path from V_1 to V_k has the form $\langle (V_1V_2)_\leftrightarrow, (V_2V_3)_\leftrightarrow, \ldots, (V_{k-2}V_{k-1})_\leftrightarrow, (V_{k-1}V_k)_\leftrightarrow \rangle$. For $V \in \vec{V}$, define the *district* (Richardson and Spirtes, 2002) or the *c-component* (Tian and Pearl, 2002) of V as the set

$$\text{dis}_{\mathscr{G}(\vec{V},\vec{W})}(V) \equiv \{Z \in \vec{V} \mid \langle (ZV_1)_\leftrightarrow, \ldots, (V_kV)_\leftrightarrow \rangle \text{ exists in } \mathscr{G}(\vec{V},\vec{W})\}.$$

The set of districts of $\mathscr{G}(\vec{V},\vec{W})$, denoted by $\mathscr{D}(\mathscr{G}(\vec{V},\vec{W}))$ is a partition of $\vec{V}$. Specifically, elements of $\mathscr{D}(\mathscr{G}(\vec{V},\vec{W}))$ correspond to connected components in the graph derived from $\mathscr{G}(\vec{V},\vec{W})$ by dropping all vertices in $\vec{W}$ and edges adjacent to such vertices, and all directed edges. In such a graph, vertices in connected components are connected by bidirected paths only, and thus correspond precisely to districts in the original CADMG. The same definition applies to any ADMG $\mathscr{G}(\vec{V})$ to yield $\mathscr{D}(\mathscr{G}(\vec{V}))$.

Other definitions, and genealogic relations defined on DAGs in Section 2.2, generalize in a straightforward way to ADMGs and CADMGs. In particular, d-separation generalizes to *m-separation* (Richardson, 2003) in ADMGs. In an ADMG, triples of the form $\langle (AC)_\to, (CB)_\leftarrow \rangle$; $\langle (AC)_\leftrightarrow, (CB)_\leftarrow \rangle$; $\langle (AC)_\to, (CB)_\leftrightarrow \rangle$; $\langle (AC)_\leftrightarrow, (CB)_\leftrightarrow \rangle$ are called colliders, and any other kind of triple a non-collider. Given an ADMG $\mathscr{G}(\vec{V})$, disjoint vertices $A, B \in \vec{V}$, and a set $\vec{C}$, such that $A, B \notin \vec{C}$, a path from A to B is said to be *m-separated* given a set $\vec{C}$ if there exists a collider $\langle (AC), (CB) \rangle$ such that $\text{de}_{\mathscr{G}(\vec{V})}(C) \cap \vec{C} = \emptyset$ or a non-collider $\langle (AC), (CB) \rangle$ such that $C \in \vec{C}$.

5.2. *Latent Projections*

Identification theory in a causal model described by a hidden variable DAG $\mathscr{G}(\vec{V} \cup \vec{H})$, where $\vec{V}$ correspond to observed variables and $\vec{H}$ to hidden variables, is often described in terms of the ADMG $\mathscr{G}(\vec{V})$ constructed from $\mathscr{G}(\vec{V} \cup \vec{H})$ via the latent projection operation (Verma and Pearl, 1990).

Given a DAG $\mathscr{G}(\vec{V} \cup \vec{H})$, a latent projection onto $\vec{V}$ is the following ADMG $\mathscr{G}(\vec{V})$. First, $\mathscr{G}(\vec{V})$ contains all directed edges in $\mathscr{G}(\vec{V} \cup \vec{H})$ between elements in $\vec{V}$. Second, for every pair $V_1, V_2 \in \vec{V}$, whenever a directed path from V_1 to V_2 exists in $\mathscr{G}$ such that all intermediate elements of the path are in $\vec{H}$, $\mathscr{G}(\vec{V})$ contains an edge $(V_1 V_2)_{\rightarrow}$. Finally, whenever a path from V_1 to V_2 without collider triples exists in $\mathscr{G}$, where all intermediate elements of the path are in $\vec{H}$, the first edge points to V_1, and the last edge points to V_2, $\mathscr{G}(\vec{V})$ contains an edge $(V_1 V_2)_{\leftrightarrow}$. As an example, the latent projection of a hidden variable DAG in Fig. 1 (c), where W, A, M, Y are observed, and H is hidden is an ADMG shown in Fig. 1 (d), while the latent projection of a hidden variable DAG in Fig. 2 (a), where W, A, Y are observed, and H is hidden is an ADMG shown in Fig. 2 (b). Corresponding to the ADMG $\mathscr{G}(\vec{V})$ is the class of all hidden variable DAGs $\mathscr{G}(\vec{V} \cup \vec{H})$ such that the latent projection of $\mathscr{G}(\vec{V} \cup \vec{H})$ onto $\vec{V}$ results in $\mathscr{G}(\vec{V})$.

The latent projection of a hidden variable DAG $\mathscr{G}(\vec{V} \cup \vec{H})$ where vertices in $\vec{H}$ corresponding to hidden variables are "projected out" is written as $\mathscr{G}(\vec{V})$. Similarly, the marginal distribution of $p(\vec{V} \cup \vec{H})$ where variables in $\vec{H}$ are marginalized out is written as $p(\vec{V})$. The notation for latent projections in graphs thus intentionally resembles the notation for marginalization in distributions, as the latent projection operation can be viewed as the graphical analogue of the marginalization operation.

5.3. *Targets Of Inference In Hidden Variable DAG Models*

Defining $Y(\vec{a})$ via (4), $Y(\pi, \vec{a}, \vec{a}')$ via (5), and $Y(\vec{g}_{\vec{A}})$ via (7) in a hidden variable DAG $\mathscr{G}(\vec{V} \cup \vec{H})$ can, in certain cases, be done directly on a latent projection $\mathscr{G}(\vec{V})$, and in a way that only variables in $\vec{V}$ are mentioned, without changing the meaning of the counterfactual. For $Y(\vec{a})$ the only requirement is that $Y \in \vec{V}$, and $\vec{A} \subseteq \vec{V}$. For $Y(\vec{g}_{\vec{A}})$, where $\vec{g}_{\vec{A}} = \{g_A(\vec{W}_A) \mid A \in \vec{A}\}$, in addition it must be the case that for every A, $\vec{W}_A \subseteq \vec{V}$.

Since directed paths in π may involve vertices in $\vec{H}$, defining a version of $Y(\pi, \vec{a}, \vec{a}')$ on the latent projection $\mathscr{G}(\vec{V})$ involves considering a counterfactual $Y(\tilde{\pi}, \vec{a}, \vec{a}')$, where $\tilde{\pi}$ is a set of directed paths containing only vertices in $\vec{V}$. Specifically, $\tilde{\pi}$ is the set of directed paths consisting of, for every path $\pi_i \in \pi$, the largest subpath of π_i consisting only of elements in $\vec{V}$. Every path in π yields a path in $\tilde{\pi}$, but multiple paths in π may end up yielding the same path in $\tilde{\pi}$. The counterfactual $Y(\tilde{\pi}, \vec{a}, \vec{a}')$ is well defined if for any $\tilde{\pi}_k \in \tilde{\pi}$ formed from $\pi_i \in \pi$, it is not the case that there exists $\pi_j \notin \pi$ such that the largest subpath of π_j consisting only of elements in $\vec{V}$ is $\tilde{\pi}_k$. This requirement is necessary for elements in $\tilde{\pi}$ to have a consistent value assignment.

Under conditions mentioned above, defining $Y(\vec{a})$ via (4), $Y(\tilde{\pi}, \vec{a}, \vec{a}')$ via (5), and $Y(\vec{g}_{\vec{A}})$ via (7) in $\mathscr{G}(\vec{V})$ yields the same counterfactual as if these definitions were applied in $\mathscr{G}(\vec{V} \cup \vec{H})$ to define $Y(\vec{a})$, $Y(\pi, \vec{a}, \vec{a}')$ and $Y(\vec{g}_{\vec{A}})$, respectively. These results follow from the way these counterfactuals are defined via (4), (5) and (6), and the definition of the latent projection ADMG, and are also described in (Shpitser, 2017). We will restrict attention to counterfactuals that can be defined on $\mathscr{G}(\vec{V})$ directly.

5.4. *Kernels As Generalized Conditional Distributions*

Functionals of identified causal parameters based on the g-formula described in section 4 can be viewed as (functions of) truncated versions of the DAG factorization (1) of $p(\vec{V})$. In hidden variable DAGs represented by latent projection ADMGs, the functionals for identified causal parameters can also be viewed as truncated versions of a particular factorization of $p(\vec{V})$ with respect to the latent projection ADMG $\mathscr{G}(\vec{V})$. However, pieces of this factorization are not simple functions of the observed distribution corresponding to conditional distributions. Instead, these pieces must be obtained from the observed distribution by an operator that sequentially applies a certain fixing operation. To describe this operator, we introduce a special type of distribution that represents, along with CADMGs introduced earlier, the operator's intermediate inputs and outputs.

A *kernel* $q_{\vec{V}}(\vec{V} \mid \vec{W})$ (Lauritzen, 1996) is a mapping from values $\vec{w}$ of $\vec{W}$ to normalized densities $q_{\vec{V}}(\vec{V} \mid \vec{w})$ over $\vec{V}$. For $\vec{W}' \subseteq \vec{W}$, $q_{\vec{V}}(\vec{V} \mid \vec{w}', \vec{W} \setminus \vec{W}')$ is a restriction of $q_{\vec{V}}(\vec{V} \mid \vec{W})$ to a mapping from values $\vec{w}_{\vec{W} \setminus \vec{W}'}$ of $\vec{W} \setminus \vec{W}'$ to normalized densities $q_{\vec{V}}(\vec{V} \mid \vec{w}' \cup \vec{w}_{\vec{W} \setminus \vec{W}'})$. A conditional distribution is a type of kernel, though other types of kernels are possible. For example, the identifying functional in (10) is a kernel $q(Y \mid a) \equiv \sum_W p(Y \mid a, W) p(W)$ that is not in general equal to the conditional distribution $p(Y \mid a)$. Given a subset $\vec{A} \subseteq \vec{V}$, marginalization and conditioning for kernels are defined in the usual way:

$$q_{\vec{V}}(\vec{A} \mid \vec{W}) \equiv \sum_{\vec{V} \setminus \vec{A}} q_{\vec{V}}(\vec{V} \mid \vec{W}); \qquad q_{\vec{V}}(\vec{V} \setminus \vec{A} \mid \vec{A} \cup \vec{W}) \equiv \frac{q_{\vec{V}}(\vec{V} \mid \vec{W})}{q_{\vec{V}}(\vec{A} \mid \vec{W})}.$$

CADMGs and kernels involved in identification are derived from $\mathscr{G}(\vec{V})$ and $p(\vec{V})$ by sequential application of the fixing operation, defined in (Richardson et al., 2017).

5.5. *The Fixing Operation And Reachable And Intrinsic Sets*

Given a CADMG $\mathscr{G}(\vec{V}, \vec{W})$, a vertex $V \in \vec{V}$ is said to be fixable if $\text{de}_{\mathscr{G}}(V) \cap \text{dis}_{\mathscr{G}}(V) = \{V\}$. Given a fixable vertex V in $\mathscr{G}$ define the fixing operator on graphs $\phi_V(\mathscr{G}(\vec{V}, \vec{W}))$ to be an operator that produces a CADMG $\tilde{\mathscr{G}}(\vec{V} \setminus \{V\}, \vec{W} \cup \{V\})$ which is obtained from $\mathscr{G}(\vec{V}, \vec{W})$ by removing all edges of the form $(WV)_{\rightarrow}$, $(WV)_{\leftrightarrow}$. As an example, M is fixable in the ADMG $\mathscr{G}^{(e)}$ shown in Fig. 2 (e), since $\text{de}_{\mathscr{G}^{(e)}}(M) \cap \text{dis}_{\mathscr{G}^{(e)}}(M) = \{M\}$. The CADMG $\phi_M(\mathscr{G}^{(e)})$ is shown in Fig. 2 (g).

Given a CADMG $\mathscr{G}(\vec{V}, \vec{W})$ and a kernel $q_{\vec{V}}(\vec{V} \mid \vec{W})$, and any fixable V in $\mathscr{G}$, define the fixing operator on kernels $\phi_V(q_{\vec{V}}(\vec{V} \mid \vec{W}); \mathscr{G}(\vec{V}, \vec{W}))$ to be one that produces the kernel

$$\tilde{q}_{\vec{V} \setminus \{V\}}(\vec{V} \setminus \{V\} \mid \vec{W} \cup \{V\}) \equiv \frac{q_{\vec{V}}(\vec{V} \mid \vec{W})}{q_{\vec{V}}(V \mid \text{nd}_{\mathscr{G}}(V) \cup \vec{W})}.$$

We remind the reader that the set of non-descendants of V, $\text{nd}_{\mathscr{G}}(V)$ is defined as every element that is not a descendant of V: $(\vec{V} \cup \vec{W}) \setminus \text{de}_{\mathscr{G}}(V)$. As an example, given the joint distribution

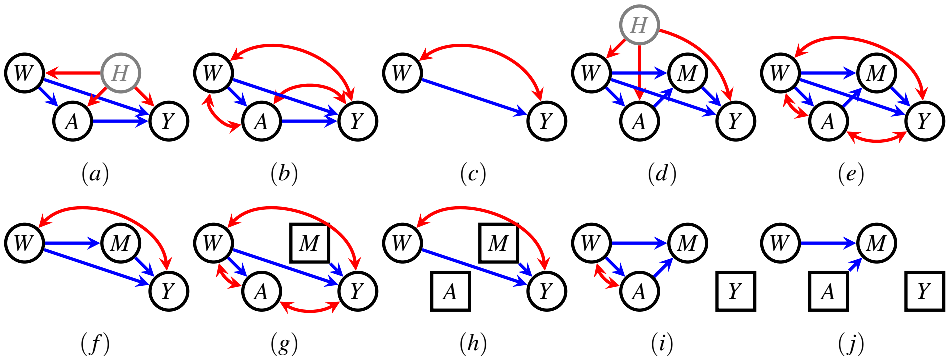

FIGURE 2. *(a) A causal DAG with an unobserved common cause of the treatment A and the outcome Y which prevents identification of $p(Y(a))$. (b) The ADMG obtained from the DAG in (a) via the latent projection operation collapsing over the unobserved variable H. (c) A subgraph of the ADMG in (b) relevant for identification of $p(Y \mid do(a))$. (d) A causal DAG with an unobserved common cause of the baseline variables W, the treatment A and the outcome Y. This DAG also contains a mediator M that "captures" all the causal influence of A on Y that is also not confounded by H. (e) The ADMG $\mathscr{G}$ obtained from the DAG in (d) via the latent projection operation collapsing over the unobserved variable H. (f) A subgraph of the ADMG in (e) relevant for identification of $p(Y \mid do(a))$. (g) The CADMG $\phi_M(\mathscr{G})$ obtained from the ADMG in (e). (h) The CADMG $\phi_{\langle M,A \rangle}(\mathscr{G}) = \phi_{\{M,A\}}(\mathscr{G})$ obtained from the ADMG in (e). (i) The CADMG $\phi_Y(\mathscr{G})$ obtained from the ADMG in (e). (j) The CADMG $\phi_{\langle Y,A \rangle}(\mathscr{G}) = \phi_{\{Y,A\}}(\mathscr{G})$ obtained from the ADMG in (e).*

$p(W,A,M,Y)$ associated with the ADMG $\mathscr{G}^{(e)}$ shown in Fig. 2 (e), we have, since M is fixable in $\mathscr{G}^{(e)}$,

$$\phi_M(p(W,A,M,Y);\mathscr{G}^{(e)}) = \frac{p(W,A,M,Y)}{p(M \mid \text{nd}_{\mathscr{G}^{(e)}}(M))} = \frac{p(W,A,M,Y)}{p(M \mid W,A)} = p(Y \mid W,A,M)p(W,A).$$

The fixing operator is applied iteratively to CADMG/kernel pairs, starting with the latent projection $\mathscr{G}(\vec{V})$ of a hidden variable DAG $\mathscr{G}(\vec{V} \cup \vec{H})$ and the observed distribution $p(\vec{V})$ which is a marginal of the true distribution $p(\vec{V} \cup \vec{H})$ that is Markov relative to $\mathscr{G}(\vec{V} \cup \vec{H})$. Not all sequences of fixing operations are allowed.

Given a vertex set $\{V_1,\ldots,V_k\} = \vec{A} \subseteq \vec{V}$ in a CADMG $\mathscr{G}(\vec{V},\vec{W})$, we will denote sequences on elements in $\vec{A}$ as $\sigma_{\vec{A}}$ or $\langle V_1,\ldots,V_k \rangle$. Given $\sigma_{\vec{A}} \equiv \langle V_1,\ldots \rangle$, where $\vec{A}$ is not the empty set, the operator $\tau(\sigma_{\vec{A}})$ yields the subsequence $\sigma_{\vec{A}}$ consisting of all but the first element V_1, in the same order as in $\sigma_{\vec{A}}$.

A sequence $\sigma_{\vec{A}}$ of the set $\vec{A} \equiv \{V_1,\ldots,V_k\}$ is said to be *fixable* in a CADMG $\mathscr{G}(\vec{V},\vec{W})$ if

(a) $\sigma_{\vec{A}} = \langle\rangle$, that is if $\vec{A}$ is the empty set, or

(b) $\sigma_{\vec{A}} \equiv \langle V_1,\ldots \rangle$ has at least one element, V_1 is fixable in $\mathscr{G}(\vec{V},\vec{W})$, and $\tau(\sigma_{\vec{A}})$ is a sequence fixable in $\phi_{V_1}(\mathscr{G}(\vec{V},\vec{W}))$.

If there exists $\sigma_{\vec{A}}$ fixable in $\mathscr{G}(\vec{V},\vec{W})$, we say $\vec{A}$ is fixable in $\mathscr{G}(\vec{V},\vec{W})$, and $\vec{R} \equiv \vec{V} \setminus \vec{A}$ is *reachable* in $\mathscr{G}(\vec{V},\vec{W})$. A reachable set $\vec{R}$ is said to be *intrinsic* if $\mathscr{G}(\vec{V})_{\vec{R}}$ contains a single district. In particular, for any reachable $\vec{R}$ (due to a fixable sequence $\sigma_{\vec{V}\setminus\vec{R}}$), all districts in $\phi_{\sigma_{\vec{V}\setminus\vec{R}}}(\mathscr{G})$ are intrinsic sets. The notions of reachable and intrinsic sets are purely graphical, hence every ADMG $\mathscr{G}(\vec{V})$ has a fixed set of reachable and intrinsic subsets of $\vec{V}$, we denote the latter by $\mathscr{I}(\mathscr{G}(\vec{V}))$.

Given a CADMG $\mathscr{G}(\vec{V},\vec{W})$, a kernel $q_{\vec{V}}(\vec{V} \mid \vec{W})$, a set $\vec{A} \subseteq \vec{V}$, and a sequence $\sigma_{\vec{A}}$ fixable in $\mathscr{G}(\vec{V},\vec{W})$, we define the fixing operator for both graphs and kernels for $\sigma_{\vec{A}}$ as follows:

If $\sigma_{\vec{A}} \equiv \langle\rangle$, that is if $\vec{A}$ is the empty set,

$$\phi_{\sigma_{\vec{A}}}(\mathscr{G}(\vec{V},\vec{W})) \equiv \mathscr{G}(\vec{V},\vec{W}) \text{ and } \phi_{\sigma_{\vec{A}}}(q_{\vec{V}}(\vec{V} \mid \vec{W});\mathscr{G}(\vec{V},\vec{W})) \equiv q_{\vec{V}}(\vec{V} \mid \vec{W}),$$

If $\sigma_{\vec{A}} \equiv \langle V_1, \ldots \rangle$,

$$\phi_{\sigma_{\vec{A}}}(\mathscr{G}(\vec{V},\vec{W})) \equiv \phi_{\tau(\sigma_{\vec{A}})}(\phi_{V_1}(\mathscr{G}(\vec{V},\vec{W}))) \text{ and}$$

$$\phi_{\sigma_{\vec{A}}}(q_{\vec{V}}(\vec{V} \mid \vec{W});\mathscr{G}(\vec{V},\vec{W})) \equiv \phi_{\tau(\sigma_{\vec{A}})}(\phi_{V_1}(q_{\vec{V}}(\vec{V} \mid \vec{W});\mathscr{G}(\vec{V},\vec{W}));\phi_{V_1}(\mathscr{G}(\vec{V},\vec{W}))).$$

For any two distinct sequences $\sigma^1_{\vec{A}} \equiv \langle V_{i_1}, \ldots, V_{i_k}\rangle$, $\sigma^2_{\vec{A}} \equiv \langle V_{j_1}, \ldots, V_{j_k}\rangle$ of vertices in $\vec{A} \subseteq \vec{V}$ fixable in $\mathscr{G}(\vec{V},\vec{W})$, it is known (Richardson et al., 2017) that $\phi_{\sigma^1_{\vec{A}}}(\mathscr{G}(\vec{V},\vec{W})) = \phi_{\sigma^2_{\vec{A}}}(\mathscr{G}(\vec{V},\vec{W}))$. That is, any fixable sequence applied to the same set of vertices in the same original CADMG yields the same final CADMG. Hence, we define the fixing operator on graphs for reachable $\vec{R}$ directly on sets as $\phi_{\vec{V}\setminus\vec{R}}(\mathscr{G}(\vec{V}))$ to mean "apply the fixing operator to elements in $\vec{V}\setminus\vec{R}$ according to some (any) fixable sequence."

As an example, the sequence $\langle M,A\rangle$ is fixable in the ADMG $\mathscr{G}^{(e)}$ shown in Fig. 2 (e). The result of applying $\phi_{\langle M,A,\rangle}(\mathscr{G}^{(e)})$ is a CADMG shown in Fig. 2 (h). Similarly, the result of applying the fixing operator according to this sequence to $p(W,A,M,Y)$ and $\mathscr{G}^{(e)}$ yields

$$\phi_{\langle M,A,\rangle}(p(W,A,M,Y);\mathscr{G}^{(e)}) = \phi_A\left(\frac{p(W,A,M,Y)}{p(M \mid W,A)};\mathscr{G}^{(g)}\right) = \frac{p(Y \mid W,A,M)p(W,A)}{\frac{p(Y|W,A,M)p(W,A)}{\sum_A p(Y|W,A,M)p(W,A)}}$$
$$= \sum_A p(Y \mid W,A,M)p(W,A).$$

5.6. *The Nested Markov Factorization*

A distribution $p(\vec{V})$ is said to nested Markov factorize with respect to, or be nested Markov relative to $\mathscr{G}(\vec{V})$ if there exists a set of kernels $\{\tilde{q}_{\vec{S}}(\vec{S} \mid \text{pa}_{\mathscr{G}}(\vec{S})\setminus\vec{S}) : \vec{S} \in \mathscr{I}(\mathscr{G}(\vec{V}))\}$ such that for every reachable $\vec{R}$ and *any* corresponding fixable sequence $\sigma_{\vec{V}\setminus\vec{R}}$,

$$\phi_{\sigma_{\vec{V}\setminus\vec{R}}}(p(\vec{V});\mathscr{G}(\vec{V})) = \prod_{\vec{D}\in\mathscr{D}(\phi_{\vec{V}\setminus\vec{R}}(\mathscr{G}(\vec{V})))} \tilde{q}_{\vec{D}}(\vec{D} \mid \text{pa}_{\mathscr{G}}(\vec{V})(\vec{D})\setminus\vec{D}).$$

In words, this says that $p(\vec{V})$ nested Markov factorizes with respect to an ADMG $\mathscr{G}(\vec{V})$ if

(a) there exists a set of intrinsic kernels corresponding to intrinsic sets in $\mathscr{G}(\vec{V})$, such that

(b) every kernel corresponding to a reachable set $\vec{R}$, obtained by applying the fixing operator to $p(\vec{V})$ and $\mathscr{G}(\vec{V})$ using the fixable sequence $\sigma_{\vec{V}\setminus\vec{R}}$,

(c) factorizes into a product of intrinsic kernels corresponding to the districts of the CADMG corresponding to $\vec{R}$,

(d) where this CADMG is obtained by applying the fixing operator to the graph $\mathscr{G}(\vec{V})$ and the set $\vec{V}\setminus\vec{R}$.

The factorization is *nested* because it describes the factorization for many kernels corresponding to many reachable sets, some "nested" within others. In particular, the set $\vec{V}$ is (trivially) reachable in $\mathscr{G}(\vec{V})$, so all other reachable sets are "nested" within $\vec{V}$. Since $\vec{V}$ is reachable, the nested factorization asserts that the original distribution $p(\vec{V})$ factorizes according to districts in $\mathscr{G}(\vec{V})$.

If $p(\vec{V})$ is nested Markov relative to the ADMG $\mathscr{G}(\vec{V})$, then for any two distinct sequences $\sigma^1_{\vec{A}} \equiv \langle V_{i_1}, \ldots, V_{i_k} \rangle$, $\sigma^2_{\vec{A}} \equiv \langle V_{j_1}, \ldots, V_{j_k} \rangle$ of vertices in $\vec{A} \subseteq \vec{V}$ fixable in $\mathscr{G}(\vec{V})$, it is known (Richardson et al., 2017) that $\phi_{\sigma^1_{\vec{A}}}(p(\vec{V});\mathscr{G}(\vec{V})) = \phi_{\sigma^2_{\vec{A}}}(p(\vec{V});\mathscr{G}(\vec{V}))$. In other words, if $p(\vec{V})$ is in the nested Markov model, then any distinct fixable sequences on the same set applied to $p(\vec{V})$ and $\mathscr{G}(\vec{V})$ lead to the same kernel. This result justifies restating the fixing operation on kernels in terms of sets as well. For any reachable $\vec{R}$, we write $\phi_{\vec{V}\setminus\vec{R}}(p(\vec{V});\mathscr{G}(\vec{V}))$ to mean "the kernel obtained from applying the fixing operator to $p(\vec{V})$ and $\mathscr{G}(\vec{V})$ according to some (any) fixable sequence."

In addition, it can be shown (Richardson et al., 2017) that the set of intrinsic kernels is, in fact $\{\phi_{\vec{V}\setminus\vec{S}}(p(\vec{V});\mathscr{G}(\vec{V})) : \vec{S} \in \mathscr{I}(\mathscr{G}(\vec{V}))\}$, and the nested Markov factorization can be rephrased as:

$$\phi_{\vec{V}\setminus\vec{R}}(p(\vec{V});\mathscr{G}(\vec{V})) = \prod_{\vec{D}\in\mathscr{D}(\phi_{\vec{V}\setminus\vec{D}}(\mathscr{G}(\vec{V}))} \phi_{\vec{V}\setminus\vec{D}}(p(\vec{V});\mathscr{G}(\vec{V})),$$

for every $\vec{R}$ reachable in $\mathscr{G}(\vec{V})$.

5.7. *The ID Algorithm*

Given a hidden variable DAG $\mathscr{G}(\vec{V} \cup \vec{H})$ representing a causal model, assume we are interested in identification of the interventional distribution $p(\vec{Y} \mid \text{do}(\vec{a}))$, for arbitrary disjoint subsets $\vec{A}, \vec{Y}$ of $\vec{V}$, from the observed distribution $p(\vec{V})$. This identification problem was given a general solution in terms of a recursive algorithm called the ID algorithm (Tian and Pearl, 2002) which was shown to be complete in (Shpitser and Pearl, 2006b; Huang and Valtorta, 2006). Here we give a simple one line reformulation of the ID algorithm as a truncated nested factorization.

Let $\mathscr{G}(\vec{V})$ be the latent projection of $\mathscr{G}(\vec{V} \cup \vec{H})$ onto observable vertices $\vec{V}$, and let $\vec{Y}^* \equiv \text{an}_{\mathscr{G}(\vec{V})_{\vec{V}\setminus\vec{A}}}(\vec{Y})$, be the set of ancestors of $\vec{Y}$ via only directed paths that exclude elements in $\vec{A}$ (in particular, $\vec{Y}^*$ does not contain any element of $\vec{A}$). Let $\mathscr{G}(\vec{V})_{\vec{Y}^*}$ be the induced subgraph of $\mathscr{G}(\vec{V})$ containing only elements in $\vec{Y}^*$ and edges in $\mathscr{G}(\vec{V})$ among those elements. Let $\mathscr{D}(\mathscr{G}(\vec{V})_{\vec{Y}^*})$ be the set of districts in $\mathscr{G}(\vec{V})_{\vec{Y}^*}$. If every $\vec{D} \in \mathscr{D}(\mathscr{G}(\vec{V})_{\vec{Y}^*})$ is an intrinsic set, then

$$p(\vec{Y} \mid \text{do}(\vec{a})) = \sum_{\vec{Y}^*\setminus\vec{Y}} \prod_{\vec{D}\in\mathscr{D}(\mathscr{G}(\vec{V})_{\vec{Y}^*})} \phi_{\vec{V}\setminus\vec{D}}(p(\vec{V});\mathscr{G}(\vec{V}))|_{\vec{A}=\vec{a}}. \tag{12}$$

If some $\vec{D} \in \mathscr{D}(\mathscr{G}(\vec{V})_{\vec{Y}^*})$ is not intrinsic, then $p(\vec{Y} \mid \text{do}(\vec{a}))$ is not identifiable in the causal model, meaning no other algorithm can obtain the identifying expression for $p(\vec{Y} \mid \text{do}(\vec{a}))$ without additional assumptions. In words, this says that $p(\vec{Y} \mid \text{do}(\vec{a}))$ is identifiable if and only if the induced graph $\mathscr{G}(\vec{V})_{\vec{Y}^*}$ for a certain set $\vec{Y}^*$ factorizes into districts that all correspond to intrinsic sets in the original ADMG $\mathscr{G}(\vec{V})$. $\vec{Y}^*$ may not be reachable itself, but still yield this sort of factorization. The factorization implies the interventional distribution of interest may be obtained from the product of the corresponding intrinsic kernels (which are all themselves identifiable), and possibly a marginalization operation that sums out elements in $\vec{Y}^* \setminus \vec{Y}$.

The proof of soundness of the original formulation of the ID algorithm appears in (Tian and Pearl, 2002), and of completeness in (Shpitser and Pearl, 2006b; Huang and Valtorta, 2006). The proof of soundness of the simplified version shown in (12) appears in (Richardson et al., 2017). Since preconditions for the application of (12), and all expressions within (12) itself were phrased in terms of the latent projection $\mathscr{G}(\vec{V})$, and not the original hidden variable DAG $\mathscr{G}(\vec{V} \cup \vec{H})$, all hidden variable DAGs that share the latent projection $\mathscr{G}(\vec{V})$ agree on which distributions $p(\vec{Y} \mid \text{do}(\vec{a}))$ are identifiable and which are not, and further agree on all identifying functionals. An implementation of the ID algorithm, and a number of related algorithms, exists in the `causaleffect` package in the R programming language.

We now illustrate how (12) is applied with two examples. The first example is the causal model corresponding to the DAG in Fig. 2 (a). This DAG contains an unobserved common cause of A and Y, which will prevent identification of $p(Y \mid \text{do}(a))$, as we now show. The latent projection of this graph is the ADMG $\mathscr{G}(\{W,A,Y\})$ shown in Fig. 2 (b). Here $\vec{Y}^* = \text{an}_{\mathscr{G}_{\{Y,W\}}}(Y) = \{Y,W\}$, with $\mathscr{G}_{\vec{Y}^*}$ shown in Fig. 2 (c). Then $\mathscr{D}(\mathscr{G}_{\vec{Y}^*}) = \{\{W,Y\}\}$, and the set $\vec{V} \setminus \{W,Y\} = \{A\}$ is not fixable, since Y is a descendant of A and lies in the district of A. Thus $p(Y \mid \text{do}(a)) = p(Y(a))$ is not identified in the causal model represented by Fig. 2 (a).

Next, consider the causal model corresponding to the DAG in Fig. 2 (d). Like in Fig. 2 (a), there is a common cause of A and Y. However, in Fig. 2 (d) there is, in addition, a mediator variable M which lies on a causal pathway from A to Y, indicated by the presence of a directed path $A \to M \to Y$. In fact, this is the only directed path from A to Y, meaning that M captures or mediates all of the causal influence of A on Y. Finally, M is not a child of H, meaning it is determined entirely by W and A and remains unconfounded by H, unlike A and Y. The presence of this type of mediator allows $p(Y \mid \text{do}(a))$ to be identified, as we now show.

The latent projection of Fig. 2 (d) is the ADMG $\mathscr{G}(\{W,A,M,Y\})$ shown in Fig. 2 (e). Here $\vec{Y}^* = \text{an}_{\mathscr{G}_{\{Y,W,M\}}}(Y) = \{Y,M,W\}$, with $\mathscr{G}_{\vec{Y}^*}$ shown in Fig. 2 (f). It's easy to verify that $\mathscr{D}(\mathscr{G}_{\vec{Y}^*}) = \{\{Y,W\},\{M\}\}$. In this case, the sets $\vec{V} \setminus \{Y,W\} = \{A,M\}$ and $\vec{V} \setminus \{M\} = \{Y,W,A\}$ are fixable. We first consider $\{A,M\}$. M is fixable in Fig. 2 (e), which yields the CADMG in Fig. 2 (g), with the corresponding kernel

$$q_{\{W,A,Y\}}(W,A,Y \mid M) = \frac{p(W,A,M,Y)}{p(M \mid A,W)} = p(Y \mid M,A,W)p(A,W).$$

In this CADMG, A becomes fixable (it was not fixable in Fig. 2 (e)), yielding the CADMG in Fig. 2 (h), with the corresponding kernel

$$q_{\{W,Y\}}(W,Y \mid A,M) = \frac{q_{\{W,A,Y\}}(W,A,Y \mid M)}{q_{\{W,A,Y\}}(A \mid W,Y,M)} = \sum_A p(Y \mid M,A,W,)p(A,W).$$

Similarly, the set $\{Y,W,A\}$ is also fixable. First, Y is fixable in Fig. 2 (e), yielding the CADMG in Fig. 2 (i), with the corresponding kernel

$$q_{\{W,A,M\}}(W,A,M \mid Y) = \frac{p(W,A,M,Y)}{p(Y \mid W,A,M)} = p(W,A,M).$$

Next, A becomes fixable in Fig. 2 (i), yielding the CADMG in Fig. 2 (j), with the corresponding kernel

$$q_{\{W,M\}}(W,M \mid A,Y) = \frac{q_{\{W,A,M\}}(W,A,M \mid Y)}{q_{\{W,A,M\}}(A \mid W,Y)} = p(M \mid A,W)p(W).$$

Finally, W is fixable in Fig. 2 (j), yielding a CADMG obtained from Fig. 2 (j) by drawing W as a square, with the corresponding kernel

$$q_{\{M\}}(M \mid W,A,Y) = \frac{q_{\{W,M\}}(W,M \mid A,Y)}{q_{\{W,M\}}(W \mid A,Y)} = p(M \mid A,W).$$

Combining these two kernels into (12), where $\vec{Y} \setminus \vec{Y}^* = \{W,M\}$, and evaluating $q_{\{M\}}(M \mid W,A,Y)$ at $A = a$ yields

$$\begin{aligned} p(Y \mid \mathrm{do}(a)) &= \sum_{W,M} \phi^{a}_{\{Y,A,W\}}(p(W,A,M,Y);\mathscr{G})\phi_{\{A,M\}}(p(W,A,M,Y);\mathscr{G}) \\ &= \sum_{W,M} q_{\{M\}}(M \mid W,a,Y)q_{\{W,Y\}}(W,Y \mid A,M) \\ &= \sum_{W,M} p(M \mid a,W)\left(\sum_{A} p(Y \mid M,A,W)p(A,W)\right). \end{aligned}$$

This is known as the front-door formula (Pearl, 2009).

Expressions obtained from (12) may become quite complicated in arbitrary ADMGs. However, the ID algorithm as expressed in (12) may be still be viewed as a kind of "nested g-formula" appropriate to the hidden variable DAG setting, in the following sense. Just as was the case in (8), no term in (12) is a density over any element in $\vec{A}$. This is because no variable in $\vec{A}$ is an element of any $\vec{D} \in \mathscr{D}(\mathscr{G}(\vec{V})_{\vec{Y}^*})$, which means all elements in $\vec{A}$ are fixed in every term in (12). In fact, removing terms of the form $p(A \mid \mathrm{pa}_{\mathscr{G}}(A))$ from the DAG factorization corresponds to fixing in DAGs without hidden variables.

5.8. *Path-Specific Effects*

Path-specific effects in DAGs with all variables observed were not always identified due to the presence of recanting witnesses. In hidden variable DAGs, an additional type of graphical structure may also prevent identification.

Given a latent projection ADMG $\mathscr{G}(\vec{V})$, fix disjoint sets $\vec{A},\vec{Y}$ and a set of proper causal paths π for $\vec{A}$ and $\vec{Y}$, where each $A \in \vec{A}$ is the origin of at least one path in π. Let $\vec{Y}^* \equiv \mathrm{an}_{\mathscr{G}_{\vec{V}\setminus\vec{A}}}(\vec{Y})$. Then $\vec{D} \in \mathscr{D}(\mathscr{G}(\vec{V})_{\vec{Y}^*})$ is said to be a *recanting district* for π if there exists a path in π of the form $\langle (AD_1)_{\rightarrow},\ldots,(W_2Y_1)_{\rightarrow}\rangle$, and another proper causal path not in π of the form $\langle (AD_2)_{\rightarrow},\ldots,(W_2Y_2)_{\rightarrow}\rangle$, where $D_1,D_2 \in \vec{D}$, and $Y_1,Y_2 \in \vec{Y}$. Identification of potential outcomes $\{Y(\pi,\vec{a},\vec{a}') \mid Y \in \vec{Y}\}$ involved in a path-specific effect is characterized in DAGs with hidden variables by the absence of a recanting district, and identifiability of the effect of $\vec{A}$ on $\vec{Y}$ along *all* paths. Specifically, given disjoint vertex sets $\vec{A},\vec{Y}$ in a DAG $\mathscr{G}(\vec{V})$, and a set π of proper causal paths for $\vec{A}$ and $\vec{Y}$, the distribution $p(\{Y(\pi,\vec{a},\vec{a}') \mid Y \in \vec{Y}\})$ is identified if and only if there does *not* exist a recanting witness for π, and $p(\vec{Y} \mid \mathrm{do}(\vec{a}))$ is identified. If $p(\{Y(\pi,\vec{a},\vec{a}') \mid Y \in \vec{Y}\})$ is identified, its identifying expression is

$$p(\{Y(\pi,\vec{a},\vec{a}') \mid Y \in \vec{Y}\}) = \sum_{\vec{Y}^* \setminus \vec{Y}} \prod_{\vec{D} \in \mathscr{D}(\mathscr{G}_{\vec{Y}^*})} \phi_{\vec{V} \setminus \vec{D}}(p(\vec{V}); \mathscr{G}(\vec{V}))|_{\mathrm{pa}_{\mathscr{G}(\vec{V})}(\vec{D}) \cap \vec{A} = \tilde{\vec{a}}_{\mathrm{pa}_{\mathscr{G}(\vec{V})}(\vec{D}) \cap \vec{A}}}, \tag{13}$$

where $\tilde{\vec{a}}_{\mathrm{pa}_{\mathscr{G}}(\vec{D}) \cap \vec{A}}$ is defined to be $\vec{a}_{\mathrm{pa}_{\mathscr{G}}(\vec{D}) \cap \vec{A}}$ if all elements in $\mathrm{pa}_{\mathscr{G}}(\vec{D}) \cap \vec{A}$ are connected to elements in $\vec{D}$ via edges in π, defined to be $\vec{a}'_{\mathrm{pa}_{\mathscr{G}}(\vec{D}) \cap \vec{A}}$ if all elements in $\mathrm{pa}_{\mathscr{G}}(\vec{D}) \cap \vec{A}$ are connected to elements in $\vec{D}$ via edges not in π, and defined to be the empty set if $\mathrm{pa}_{\mathscr{G}}(\vec{D}) \cap \vec{A} = \emptyset$. The absence of a recanting district guarantees these three possibilities are exhaustive. Just as (12) was the appropriate nested generalization of the g-formula (8) to hidden variable DAGs, so is (13) the appropriate nested generalization of the edge g-formula (11) to hidden variable DAGs. The original version of this algorithm was described in (Shpitser, 2013), with the above reformulation based on the fixing operator found in (Shpitser and Sherman, 2018).

Consider Fig. 1 (c), where we are interested in the path-specific effect of A on Y via the path $\langle (AZ)_{\rightarrow}, (ZY)_{\rightarrow} \rangle$. The latent projection of this graph is shown in Fig. 1 (d). Here $\vec{Y}^* = \{W, M, Z, Y\}$, and $\mathscr{D}(\mathscr{G}_{\vec{Y}^*}) = \{\{W\}, \{Z\}, \{M, Y\}\}$. Note that there is no recanting district – the district containing the first post-exposure variable on the only path of interest is $\{Z\}$, and no path other than $\langle (AZ)_{\rightarrow}, (ZY)_{\rightarrow} \rangle$ has the first post-exposure variable in this district. Furthermore, $p(Y \mid \mathrm{do}(a))$ is identifiable. Thus, the counterfactual corresponding to the path-specific effect is identified:

$$\begin{aligned} p(Y(\pi, a, a')) &= \sum_{W,Z,M} \left(\phi_{\{W,A,Z\}}(p; \mathscr{G}^{(d)})|_{A=a'} \right) \cdot \left(\phi_{\{W,A,M,Y\}}(p; \mathscr{G}^{(d)})|_{A=a} \right) \cdot \left(\phi_{\{A,M,Z,Y\}}(p; \mathscr{G}^{(d)}) \right) \\ &= \sum_{W,Z,M} \left(p(Y \mid a', M, Z, W) p(M \mid a', W) \right) p(Z \mid a, M, W) p(W), \end{aligned}$$

where $\mathscr{G}^{(d)}$ is the graph shown in Fig. 1 (d).

On the other hand, if we were interested in the path-specific effect of A on Y along paths $\pi = \{\langle (AZ)_{\rightarrow}, (ZY)_{\rightarrow} \rangle; \langle (AY)_{\rightarrow} \rangle\}$, this path-specific effect is not identified. This is because the path $\langle (AM)_{\rightarrow}, (MY)_{\rightarrow} \rangle$ is not in π but has $(AM)_{\rightarrow}$ as the first edge, while $\langle (AY)_{\rightarrow} \rangle$ is a path in π. M and Y share a district in $\mathscr{G}(\vec{V})_{\vec{Y}^*}$, where $\vec{Y}^* = \{W, M, Z, Y\}$. This implies $\{M, Y\}$ is a recanting district, and will prevent identification of $p(Y(\pi, a, a'))$.

5.9. *Responses To Dynamic Treatment Regimes*

An adaption of the ID algorithm for identification of distributions over responses to dynamic treatment regimes of the form $p(\{Y(\vec{g}_{\vec{A}}) \mid Y \in \vec{Y}\})$ in causal models represented by a hidden variable DAG $\mathscr{G}(\vec{H} \cup \vec{V})$ was given in (Tian, 2008).

As before, we rephrase this algorithm in terms of the fixing operation, CADMGs and kernels. Given a latent projection $\mathscr{G}(\vec{V})$ of $\mathscr{G}(\vec{H} \cup \vec{V})$, define the graph $\mathscr{G}(\vec{V})_{\vec{g}_{\vec{A}}}$ to be an ADMG obtained from $\mathscr{G}(\vec{V})$ by removing all edges pointing into $\vec{A}$ and adding a directed edge $W \rightarrow A$ for any $W \in \vec{W}_A$. Define $\vec{Y}^* \equiv \mathrm{an}_{\mathscr{G}(\vec{V})_{\vec{g}_{\vec{A}}}}(\vec{Y}) \setminus \vec{A}$. Then $p(\{Y(\vec{g}_{\vec{A}}) \mid Y \in \vec{Y}\})$ is identified if $p(\vec{Y}^* \mid \mathrm{do}(\vec{a}))$ is identified. Moreover, the identification formula is

$$p(\{Y(\vec{g}_{\vec{A}}) \mid Y \in \vec{Y}\}) = \sum_{(\vec{Y}^* \cup \vec{A}) \setminus \vec{Y}} \prod_{\vec{D} \in \mathcal{D}(\mathcal{G}(\vec{V})_{\mathbf{Y}^*})} \phi_{\vec{V} \setminus \vec{D}}(p(\vec{V}); \mathcal{G}(\vec{V}))|_{\mathrm{pa}_{\mathcal{G}(\vec{V})}(\vec{D}) \cap \vec{A} = \tilde{\vec{a}}_{\mathrm{pa}_{\mathcal{G}(\vec{V})}(\vec{D}) \cap \vec{A}}},$$

where for every $\vec{D}$, $\tilde{\vec{a}}_{\mathrm{pa}_{\mathcal{G}(\vec{V})}(\vec{D}) \cap \vec{A}}$ is defined to be $\{A = g_A(\vec{W}_A) \mid A \in \mathrm{pa}_{\mathcal{G}(\vec{V})}(\vec{D}) \cap \vec{A}\}$ if $\mathrm{pa}_{\mathcal{G}(\vec{V})}(\vec{D}) \cap \vec{A}$ is not empty, and is defined to be the empty set otherwise. The sum over $\vec{A}$ is vacuous if $\vec{g}_{\vec{A}}$ is a set of deterministic policies, since in this case there is no variation in values of $\vec{A}$ in any $Y(\vec{g}_{\vec{A}})$. This algorithm was shown to be complete in (Shpitser and Sherman, 2018).

As an example, in Fig. 1 (c), if $\vec{g}_{\vec{A}} = \{g_A(W), g_Z(M,A,W)\}$, $\mathcal{G}(\vec{V})$ is shown in Fig. 1 (d), and $\mathcal{G}(\vec{V})_{\vec{g}_{\vec{A}}}$ is the same graph as $\mathcal{G}(\vec{V})$. Since $p(Y,M,W \mid \mathrm{do}(a,z))$ is identified as

$$\phi_{\{A,M,Z,Y\}}(p; \mathcal{G}(\vec{V}))\phi_{\{W,A,Z\}}(p; \mathcal{G}(\vec{V}))|_{A=a,Z=z} = p(W)p(Y \mid z,M,a,W)p(M \mid a,W),$$

$p(Y(\vec{g}_{\vec{A}}))$ is also identified as

$$\begin{aligned} &\sum_{W,M,A,Z} \phi_{\{A,M,Z,Y\}}(p; \mathcal{G}(\vec{V}))\phi_{\{W,A,Z\}}(p; \mathcal{G}(\vec{V}))|_{\{A=g_A(W), Z=g_Z(M,A,W)\}} \\ &= \sum_{W,M,A,Z} p(W)p(Y|Z = g_Z(M, A = g_A(W), W), M, A = g_A(W), W)p(M|A = g_A(W), W). \end{aligned}$$

6. Conditional Counterfactual Distributions

A common version of the identification problem is identification of conditional interventional distributions $p(\vec{Y} \mid \vec{Z}, \mathrm{do}(\vec{a})) = p(\vec{Y}(\vec{a}) \mid \vec{Z}(\vec{a}))$ which is defined as $p(\vec{Y} \cup \vec{Z} \mid \mathrm{do}(\vec{a}))/p(\vec{Z} \mid \mathrm{do}(\vec{a}))$ or $p(\vec{Y}(\vec{a}) \cup \vec{Z}(\vec{a}))/p(\vec{Z}(\vec{a}))$. These types of distributions are of interest in causal inference applications where *effect modification*, variation of causal effects within particular subpopulations, is of interest.

Characterizing identification of these distributions is possible using (12), (13) and a subset of $\vec{Z}$ obtained using a generalized version of a conditional ignorability argument, where the necessary conditional independence is read off from a particular version of a causal graph. Prior work phrased the independence needed in terms of rule 2 of do-calculus (Pearl, 2009). Here we describe a generalization of do-calculus called *potential outcome calculus* that describes identities governing distributions on potential outcomes defined via (4), based on graphs describing the Markov properties of distributions over these potential outcomes. These graphs are called single world intervention graphs (SWIGs) (Richardson and Robins, 2013). The advantage of the formulation we describe here is it generalizes in a straightforward way to counterfactuals not readily expressed by means of Pearl's do(.) operator, such as counterfactuals that describe path-specific effects (Malinsky et al., 2019).

6.1. Single World Intervention Graphs

Given a causal DAG $\mathcal{G}(\vec{V})$ and a set of variables $\vec{A}$ we wish to intervene on, we construct the SWIG $\mathcal{G}(\vec{a})$ as follows from $\mathcal{G}(\vec{V})$. Each vertex $A \in \vec{A}$ in $\mathcal{G}(\vec{V})$ is split into a random vertex A

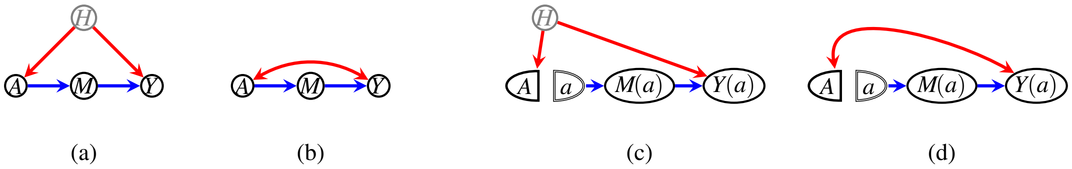

FIGURE 3. *(a) A hidden variable graph corresponding to the observed data distribution. (b) The latent projection of the graph in (a), assuming H is a hidden variable. (c) A SWIG corresponding to the world where A is set to value a, derived from the DAG in (a). (d) A latent projection version of the SWIG in (c).*

and a fixed vertex a (note the lower case). All edges with arrowheads into A in $\mathscr{G}(\vec{V})$ are inherited by the random vertex, and all directed edges out of A in $\mathscr{G}(\vec{V})$ are inherited by the fixed vertex. All other edges in $\mathscr{G}(\vec{V})$ remain in $\mathscr{G}(\vec{a})$. Finally, a vertex in $\mathscr{G}(\vec{a})$ corresponding to V in $\mathscr{G}(\vec{V})$ is labeled as a counterfactual $V(\vec{a}_V)$ with a subset $\vec{a}_V$ consisting of all elements of $\vec{a}$ with a directed path in $\mathscr{G}(\vec{a})$ to the vertex corresponding to V. As an example, given a hidden variable DAG $\mathscr{G}$ in Fig 3 (a), the SWIG $\mathscr{G}(a)$ is shown in Fig. 3 (c).

The resulting vertices $\{V(\vec{a}_V) : V \in \vec{V}\}$ correspond to the set of counterfactuals $\vec{V}(\vec{a}) \equiv V(\vec{a})$ defined by (4). The following result was proved in (Richardson and Robins, 2013):

$$p(\vec{V}(\vec{a})) = \prod_{V(\vec{a}_V) \in \vec{V}(\vec{a})} p(V(\vec{a}_V) | \text{pa}_{\mathscr{G}(\vec{a})}(V(\vec{a}_V))).$$

In other words $p(\vec{V}(\vec{a}))$ Markov factorizes with respect to the SWIG $\mathscr{G}(\vec{a})$. An important consequence of this result is that conditional independences in $p(\vec{V}(\vec{a}))$ may be directly read off from $\mathscr{G}(\vec{a})$ using d-separation. As an example, $p(A, H, M(a), Y(a))$ Markov factorizes with respect to the SWIG shown in Fig. 3 (c):

$$p(A, H, M(a), Y(a)) = p(H)p(A \mid H)p(Y(a) \mid M(a), H)p(M(a)).$$

We can therefore use d-separation in the SWIG to obtain independence statements that hold in the model. For example, the conditional ignorability constraint $(A \perp\!\!\!\perp Y(a) \mid H)$ holds by d-separation in Fig. 3 (c).

SWIGs may be constructed from latent projection ADMGs $\mathscr{G}(\vec{V})$ of a hidden variable DAG $\mathscr{G}(\vec{V} \cup \vec{H})$ by an identical "splitting" procedure, and conditional independences of a marginal counterfactual distribution $p(\vec{V}(\vec{a}))$ may be directly read off from $\mathscr{G}(\vec{a})$ using m-separation, by a simple corollary of the above result. As an example, the SWIG $\mathscr{G}(a)$ in Fig. 3 (d) is obtained from the latent projection in Fig. 3 (b) obtained from Fig. 3 (a). In this SWIG, $A \perp\!\!\!\perp M(a)$ can be obtained by m-separation.

6.2. *The Potential Outcomes Calculus*

Potential outcomes calculus is the following three identities with preconditions given by d-separation (or m-separation) on SWIGs. Fix disjoint subsets $\vec{Y}, \vec{Z}, \vec{W}, \vec{X}$ of $\vec{V}$. Then

1 $p(\vec{Y}(\vec{x}) | \vec{Z}(\vec{x}), \vec{W}(\vec{x})) = p(\vec{Y}(\vec{x}) | \vec{Z}(\vec{x}))$ if $(\vec{Y}(\vec{x}) \perp\!\!\!\perp \vec{Z}(\vec{x}) \mid \vec{W}(\vec{x}))$ in $\mathscr{G}(\vec{x})$.

2 $p(\vec{Y}(\vec{x},\vec{z})) = p(\vec{Y}(\vec{x})|\vec{W}(\vec{x}),\vec{Z}(\vec{x}) = \vec{z})$ if $\vec{Y}(\vec{x},\vec{z}) \perp\!\!\!\perp \vec{Z}(\vec{x},\vec{z}) \mid \vec{W}(\vec{x},\vec{z})$ in $\mathscr{G}(\vec{x},\vec{z})$.

3 $p(\vec{Y}(\vec{x},\vec{z})) = p(\vec{Y}(\vec{x}))$ if $\vec{Y}(\vec{x},\vec{z}) \perp\!\!\!\perp \vec{z}$ in $\mathscr{G}(\vec{x},\vec{z})$.

Rule 1 encodes the fact that conditional independences in a counterfactual distribution $p(\vec{V}(\vec{x}))$ may be read off by d-separation or m-separation in the appropriate SWIG $\mathscr{G}(\vec{x})$. Rule 2 is a kind of generalized conditional ignorability that governs where interventions on $\vec{z}$ are equivalent with conditioning on $\vec{z}$, for a particular set of variables given that other variables were either intervened on ($\vec{X}$), or conditioned on ($\vec{W}$). Classical conditional ignorability assumption is often assumed directly, whereas here conditional ignorability type statements are read off by m-separation from the appropriate SWIG $\mathscr{G}(\vec{x},\vec{z})$.

Rules 1 and 2 are straightforward analogues of Pearl's do-calculus rules, rephrased in terms of SWIGs and potential outcomes. Rule 3 we state here is far simpler than Pearl's rule 3, and states that a counterfactual $\vec{Y}(\vec{x},\vec{z})$ does not depend on $\vec{z}$ if it is the case that the corresponding set of fixed vertices in the SWIG $\mathscr{G}(\vec{x},\vec{z})$ are d-separated (or m-separated) from $\vec{Y}(\vec{x},\vec{z})$, meaning that there is no directed path from the former to the latter. This rule simply encodes the fact that recursive substitution (4) for any $Y(\vec{a})$ may yield a counterfactual that does not depend on some values in $\vec{a}$. This may occur if the corresponding variables are in the future relative to Y, or due to some exclusion restriction in the causal model, as may happen with instrumental variable models. Importantly, it was shown in (Malinsky et al., 2019) that simplifying Pearl's rule 3 in this way is *without loss of generality*: all do-calculus derivations are still possible to derive with the 3 rules given here, including the simpler rule 3.

Rule 1 may be termed the "interventional global Markov property," rule 2 may be termed "generalized conditional ignorability," and rule 3 may be termed "causal irrelevance." Rule 2 is of particular relevance to identification theory we present below.

6.3. *Conditional Causal Effects*

Given rule 2 of potential outcomes calculus, identification of distributions of the form $p(\vec{Y}(\vec{a}) \mid \vec{Z}(\vec{a}) = \vec{z})$ is quite simple to characterize. Let $\vec{W}$ be any maximal subset of $\vec{Z}$ such that for any $\vec{z}$, $p(\vec{Y}(\vec{a}) \mid \vec{Z}(\vec{a}) = \vec{z}) = p(\vec{Y}(\vec{a},\vec{w}) \mid \{Z(\vec{a},\vec{w}) : Z \in \vec{Z} \setminus \vec{W}\})$. Such a set is the unique largest set to which rule 2 applies.

Given this set, the distribution $p(\vec{Y}(\vec{a}) \mid \vec{Z}(\vec{a}) = \vec{z})$ is identifiable if and only if $p(\vec{Y}(\vec{a}), \{Z(\vec{a},\vec{w}) : Z \in \vec{Z} \setminus \vec{W}\})$ is identifiable. Moreover, the identification formula is given by

$$p(\vec{Y}(\vec{a}) \mid \vec{Z}(\vec{a}) = \vec{z}) = \frac{p(\vec{Y}(\vec{a}), \{Z(\vec{a},\vec{w}) = z : Z \in \vec{Z} \setminus \vec{W}\})}{p(\{Z(\vec{a},\vec{w}) = z : Z \in \vec{Z} \setminus \vec{W}\})},$$

where $p(\vec{Y}(\vec{a}), \{Z(\vec{a},\vec{w}) = z : Z \in \vec{Z} \setminus \vec{W}\})$ is identified via (12), and $p(\{Z(\vec{a},\vec{w}) = z : Z \in \vec{Z} \setminus \vec{W}\})$ is obtained from $p(\vec{Y}(\vec{a}), \{Z(\vec{a},\vec{w}) = z : Z \in \vec{Z} \setminus \vec{W}\})$ via marginalization (Shpitser and Pearl, 2006a).

As an example, in $\mathscr{G}^{(e)}$ shown in Fig. 2 (e), if the conditional interventional distribution $p(Y(a) \mid W(a) = w)$ is of interest, we conclude that $p(Y(a) \mid W(a) = w) = p(Y(a) \mid W = w) \neq$

$p(Y(a,w))$, since W and $Y(a,w)$ are not m-separated in the SWIG $\mathscr{G}(a,w)$ derived from $\mathscr{G}^{(e)}$. Thus, to identify $p(Y(a) \mid W(a) = w)$ we first identify $p(Y(a), W(a))$ as

$$\sum_M p(M \mid a, W) \left(\sum_A p(Y \mid M, A, W) p(A, W) \right),$$

which yields

$$p(Y(a) \mid W(a) = w) = \frac{Y(a), W(a) = w}{p(W(a)}|_{W(a)=w} = \sum_M p(M \mid a, w) \left(\sum_A p(Y \mid M, A, w) p(A \mid w) \right)$$

An extension of these results that gives a complete identification formula for conditional distributions associated with path-specific effects, using rule 2 of the potential outcomes calculus is described in (Malinsky et al., 2019).

7. Conclusion

In this paper we introduced a number of functions of potential outcome random variables used in causal inference applications, including treatment effects, direct, indirect, and path-specific effects, and responses to dynamic treatment regimes. We described a simple characterization of identifiability of these targets in fully observed causal models of a directed acyclic graph based on variations of the g-formula (Robins, 1986). Finally, we gave a simplified description of identification algorithms for these targets in causal models of a directed acyclic graph where some variables are unobserved, which first appeared in (Tian and Pearl, 2002; Tian, 2008; Shpitser and Pearl, 2006a; Shpitser, 2013). This description was based on a truncated nested Markov factorization, which is itself based on conditional mixed graphs, kernels, and the fixing operator described in (Richardson et al., 2017). The identification algorithms described are known to be complete for treatment effects, conditional treatment effects, and path-specific effects, and responses to dynamic treatment regimes.

7.1. Acknowledgements

Definitions, some examples, and some identification results that appear in this review manuscript also appeared in (Shpitser, 2017).

References

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*, volume 19, pages 357–363. Morgan Kaufmann, San Francisco.

Chakraborty, B. and Moodie, E. E. M. (2013). *Statistical Methods for Dynamic Treatment Regimes (Reinforcement Learning, Causal Inference, and Personalized Medicine)*. Springer, New York.

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

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1–12.

Huang, Y. and Valtorta, M. (2006). Pearl's calculus of interventions is complete. In *Twenty Second Conference On Uncertainty in Artificial Intelligence*.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford, U.K.: Clarendon.

Malinsky, D., Shpitser, I., and Richardson, T. S. (2019). A potential outcomes calculus for identifying conditional path-specific effects. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*.

Miles, C., Shpitser, I., Kanki, P., Melone, S., and Tchetgen Tchetgen, E. J. (2017). Quantifying an adherence path-specific effect of antiretroviral therapy in the nigeria pepfar program. *Journal of the American Statistical Association*.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences agaricales: Essay des principle. excerpts reprinted (1990) in English. *Statistical Science*, 5:463–472.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan and Kaufmann, San Mateo.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 411–420. Morgan Kaufmann, San Francisco.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition.

Pearl, J. (2011). The causal mediation formula – a guide to the assessment of pathways and mechanisms. Technical Report R-379, Cognitive Systems Laboratory, University of California, Los Angeles.

Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *Annals of Statistics*, 30:962–1030.

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

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. (2017). Nested Markov properties for acyclic directed mixed graphs. Working paper.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *preprint: http://www.csss.washington.edu/Papers/wp128.pdf*.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512.

Robins, J. M. (1999a). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*. NY: Springer-Verlag.

Robins, J. M. (1999b). Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In Glymour, C. and Cooper, G., editors, *Computation, Causation, and Discovery*, pages 349 – 405. Menlo Park, CA, CAmbridge, MA: AAAI Press/The MIT Press.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability of direct and indirect effects. *Epidemiology*, 3:143–155.

Robins, J. M. and Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects. *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*.

Rubin, D. B. (1976). Causal inference and missing data (with discussion). *Biometrika*, 63:581–592.

Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science (Rumelhart special issue)*, 37:1011–1035.

Shpitser, I. (2015). Segregated graphs and marginals of chain graph models. In *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc.

Shpitser, I. (2017). Identification in graphical causal models. In *Handbook of Graphical Models*.

Shpitser, I. and Pearl, J. (2006a). Identification of conditional interventional distributions. In *Proceedings of the Twenty Second Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 437–444. AUAI Press, Corvallis, Oregon.

Shpitser, I. and Pearl, J. (2006b). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*. AAAI Press, Palo Alto.

Shpitser, I. and Sherman, E. (2018). Identification of personalized effects associated with causal pathways. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*.

Shpitser, I. and Tchetgen Tchetgen, E. J. (2016). Causal inference with a graphical hierarchy of interventions. *Annals of Statistics*, 44(6):2433–2466.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. Springer Verlag, New York, 2 edition.

Tian, J. (2008). Identifying dynamic sequential plans. In *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 554–561, Corvallis, Oregon. AUAI Press.

Tian, J. and Pearl, J. (2002). On the testable implications of causal models with hidden variables. In *Proceedings of* *the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-02)*, volume 18, pages 519–527. AUAI Press, Corvallis, Oregon.

Verma, T. S. and Pearl, J. (1990). Equivalence and synthesis of causal models. Technical Report R-150, Department of Computer Science, University of California, Los Angeles.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20:557–585.