Identification of Personalized Effects Associated With Causal Pathways

Ilya Shpitser
Department of Computer Science
Johns Hopkins University
Baltimore, MD

Eli Sherman
Department of Computer Science
Johns Hopkins University
Baltimore, MD

Abstract

Unlike classical causal inference, where the goal is to estimate average causal effects within a *population*, in settings such as personalized medicine, the goal is to map a unit's characteristics to a treatment tailored to maximize the expected outcome for that unit. Obtaining high-quality mappings of this type is the goal of the dynamic treatment regime literature. In healthcare settings, optimizing policies with respect to a particular causal pathway is often of interest as well. In the context of average treatment effects, estimation of effects associated with causal pathways is considered in the mediation analysis literature.

In this paper, we combine mediation analysis and dynamic treatment regime ideas and consider how unit characteristics may be used to tailor a treatment strategy that maximizes an effect along specified sets of causal pathways. In particular, we define counterfactual responses to such policies, give a general identification algorithm for these counterfactuals, and prove completeness of the algorithm for unrestricted policies. A corollary of our results is that the identification algorithm for responses to policies given in [16] is complete for arbitrary policies.

1 INTRODUCTION

Establishing causal relationships between actions and outcomes is fundamental to rational decision-making. The gold standard for establishing causal relationships is the randomized controlled trial (RCT), which may be used to establish average causal effects within a population. Causal inference is a branch of statistics that seeks to predict effects of RCTs from observational data, where treatment assignment is not randomized. Such data is often gathered from observational studies, surveys given to patients during follow up, and in-hospital electronic medical records.

While average treatment effects reported from implemented RCTs, or hypothetical RCTs emulated by causal inference methods using observational data establish whether a particular action is helpful *on average*, optimal decision making must tailor decisions to specific situations. In the context of causal inference this involves finding a map between characteristics of an experimental unit, such as baseline features, to an action that optimizes some outcome for that unit. Methods for finding such maps are studied in the dynamic treatment regime literature [3], and in off-policy reinforcement learning [2].

If an action is known to have a beneficial effect on some outcome, it is often desirable to understand the causal mechanism behind this effect. A popular type of mechanism analysis is *mediation analysis*, which seeks to decompose average treatment effects into direct and indirect components, or more generally into components associated with specific causal pathways. These components of the average causal effect are known as direct, indirect, and path-specific effects, and are also defined as population averages [1, 8, 12].

In this paper, we define counterfactual outcomes necessary to personalize effects associated with causal pathways, give an algorithm for non-parametric identification of these outcomes and prove that it is complete for arbitrary policies. We consider estimation methods for identified outcomes of this type in a companion paper [7].

Why Personalize Effects Along Causal Pathways?

It often makes sense to structure decision-making such that the *overall* effect of an action on the outcome is maximized for a given unit. However, in some cases it is ap-

propriate to choose an action such that only a part of the effect of an action on the outcome is maximized. Consider management of HIV patients' care. Since HIV is a chronic disease, care for HIV patients involves designing a long-term treatment plan to minimize the chance of viral failure (an undesirable outcome). In designing such a plan, an important choice is when to initiate primary therapy, and when to switch to a second line therapy. Initiating or switching too early risks unneeded side effects and "wasting" treatment efficacy, while initiating or switching too late risks viral failure [4].

In the context of HIV, however, *treatment adherence* is an important component of the overall effect of the drug on the outcome. Patients who do not take prescribed doses compromise the efficacy of the drug, and different drugs may have different levels of adherence. Thus, for HIV patients, the overall effect of the drug can be viewed as a combination of the chemical effect and the adherence effect [6]. Therefore, choosing an action that maximizes the overall effect of HIV treatment on viral failure entangles these two very different causal mechanisms. One approach to tailoring treatments to patients in a way that disentangles these mechanisms is to find a policy that optimizes a part of the effect, say the chemical (direct) effect of the drug, while hypothetically keeping the adherence levels to some reference level. Finding such a policy yields information on how best to assign drugs to maximize their chemical efficacy in settings where adherence levels can be controlled to that of a reference treatment – even if the only data available is one where patients have differential adherence.

2 PRELIMINARIES

We proceed as follows. We first give graph theoretic preliminaries, and define graphical causal models that equate counterfactual responses to interventions (setting variables to values, contrary to fact) with truncated factorizations of the observed data distribution [11]. Next, we describe the more general *edge intervention* that sets variables to different values for different outgoing edges in a graph. Edge interventions are used to formulate direct, indirect, and path-specific effects in mediation analysis. Then, we define counterfactual responses to policies that set variables not to *constant* values but to values that potentially depend on other sets of variables. Extending these notions, we describe counterfactuals that generalize both responses to edge interventions, and responses to policies, namely responses to *edge-specific policies*. We briefly describe identification theory for these counterfactuals in causal models with no hidden variables, and note this theory is based on variations of a truncated factorization known as the g-formula [11].

We next consider identification theory for all counterfactuals we described in hidden variable causal models. This theory is more complex, and is based on the ID algorithm [14, 17]. We rephrase the algorithm and its necessary variations in a single line formula based on the fixing operator described in [10]. This reformulation allows us to express any functional corresponding to a counterfactual distribution identifiable in a hidden variable causal model as a single truncated factorization formula, just as identifiable counterfactual distributions in fully observed models are expressed via the g-formula. Finally, we describe a completeness result for the identification algorithm for responses to unrestricted edge-specific policies in hidden variable causal models.

While our primary contributions lie in the presentation of counterfactuals and identification theory for edge-specific policies, we include some discussion of prior theory to build up to our result, and show how identification theory of edge-specific policies generalizes identification theory for edge-specific effects and policy interventions.

Graph Theory

We will define statistical and causal models as sets of distributions defined by restrictions associated with graphs. We will use vertices and variables interchangeably – capital letters for a vertex or variable (V), bold capital letter for a set ($\mathbf{V}$), lowercase letters for values (v), and bold lowercase letters for sets of values ($\mathbf{v}$). By convention, each graph is defined on a vertex set $\mathbf{V}$.

For a set of values $\mathbf{a}$ of $\mathbf{A}$, and a subset $\mathbf{A}^\dagger \subseteq \mathbf{A}$, define $\mathbf{a}_{\mathbf{A}^\dagger}$ to be a restriction of $\mathbf{a}$ to elements in $\mathbf{A}^\dagger$. The state space of A will be denoted by $\mathfrak{X}_A$, and the (Cartesian product) state space of $\mathbf{A}$ will be denoted by $\mathfrak{X}_{\mathbf{A}}$.

For a graph mixed graph $\mathcal{G}$ with directed and bidirected edges, and any $V \in \mathbf{V}$, we define the following genealogic sets: parents, children, ancestors, descendants, and districts as: $\mathrm{pa}_{\mathcal{G}}(V) \equiv \{W \in \mathbf{V} \mid W \to V\}$, $\mathrm{ch}_{\mathcal{G}}(V) \equiv \{W \in \mathbf{V} \mid V \to W\}$, $\mathrm{an}_{\mathcal{G}}(V) \equiv \{W \in \mathbf{V} \mid W \to \ldots \to V\}$, $\mathrm{de}_{\mathcal{G}}(V) \equiv \{W \in \mathbf{V} \mid V \to \ldots \to W\}$, $\mathrm{dis}_{\mathcal{G}}(V) \equiv \{W \in \mathbf{V} \mid V \leftrightarrow \ldots \leftrightarrow W\}$. By convention, $\mathrm{an}_{\mathcal{G}}(V) \cap \mathrm{de}_{\mathcal{G}}(V) \cap \mathrm{dis}_{\mathcal{G}}(V) = \{V\}$. These sets generalize to $\mathbf{V}^\dagger \subseteq \mathbf{V}$ disjunctively. For example, $\mathrm{pa}_{\mathcal{G}}(\mathbf{V}^\dagger) \equiv \bigcup_{V \in \mathbf{V}^\dagger} \mathrm{pa}_{\mathcal{G}}(V)$. For $\mathbf{A} \subseteq \mathbf{V}$, define $\mathrm{pa}^s_{\mathcal{G}}(\mathbf{A}) \equiv \mathrm{pa}_{\mathcal{G}}(\mathbf{A}) \setminus \mathbf{A}$, the parents of a set $\mathbf{A}$.

The non-descendants of V are denoted $\mathrm{nd}_{\mathcal{G}}(V) \equiv \mathbf{V} \setminus \mathrm{de}_{\mathcal{G}}(V)$. The set of districts forms a partition of vertices in $\mathcal{G}$ and is denoted $\mathcal{D}(\mathcal{G})$. Finally, given a graph $\mathcal{G}$ and $\mathbf{A} \subseteq \mathbf{V}$, the subgraph of $\mathcal{G}$ containing only vertices in $\mathbf{A}$ and edges between these vertices is denoted $\mathcal{G}_{\mathbf{A}}$.

Statistical And Causal Models Of A Dag

A directed acyclic graph (DAG), or Bayesian network, is a graph $\mathcal{G}$ with vertex set $\mathbf{V}$ connected by directed edges and such that there are no directed cycles in the graph (i.e. no sequences of edges and vertices $V \rightarrow \ldots W$ and edge $W \rightarrow V$). A statistical model of a DAG $\mathcal{G}$ is the set of distributions $p(\mathbf{V})$ such that $p(\mathbf{V}) = \prod_{V \in \mathbf{V}} p(V|\, \mathrm{pa}_{\mathcal{G}}(V))$. Such a $p(\mathbf{V})$ is said to be Markov relative to $\mathcal{G}$.

Causal models of a DAG are also sets of distributions, but on counterfactual random variables. Given $Y \in \mathbf{V}$ and $\mathbf{A} \subseteq \mathbf{V} \setminus \{Y\}$, a counterfactual variable, or 'potential outcome', written as $Y(\mathbf{a})$, represents the value of Y in a hypothetical situation where $\mathbf{A}$ were set to values $\mathbf{a}$ by an *intervention operation* [9]. Given a set $\mathbf{Y}$, define $\mathbf{Y}(\mathbf{a}) \equiv \{\mathbf{Y}\}(\mathbf{a}) \equiv \{Y(\mathbf{a}) \mid Y \in \mathbf{Y}\}$. The distribution $p(\mathbf{Y}(\mathbf{a}))$ is sometimes written as $p(\mathbf{Y}|\mathrm{do}(\mathbf{a}))$ [9].

Causal models of a DAG $\mathcal{G}$ consist of distributions defined on counterfactual random variables of the form $V(\mathbf{a})$ where $\mathbf{a}$ are values of $\mathrm{pa}_{\mathcal{G}}(V)$. In this paper we assume Pearl's functional model for a DAG $\mathcal{G}$ with vertices $\mathbf{V}$ which is the set containing any joint distribution over all potential outcome random variables where the sets of variables

$$\{\{V(\mathbf{a}_V) \mid \mathbf{a}_V \in \mathfrak{X}_{\mathrm{pa}_{\mathcal{G}}(V)}\} \mid V \in \mathbf{V}\}$$

are mutually independent [9]. The *atomic counterfactuals* in the above set model the relationship between $\mathrm{pa}_{\mathcal{G}}(V)$, representing direct causes of V, and V itself. From these, all other counterfactuals may be defined using recursive substitution. For any $\mathbf{A} \subseteq \mathbf{V} \setminus \{V\}$,

$$V(\mathbf{a}) \equiv V(\mathbf{a}_{\mathrm{pa}_{\mathcal{G}}(V) \cap \mathbf{A}}, \{\mathrm{pa}_{\mathcal{G}}(V) \setminus \mathbf{A}\}(\mathbf{a})). \qquad (1)$$

For example, in the DAG in Fig. 1 (a), $Y(a)$ is defined to be $Y(a, M(a, W), W)$.

A causal parameter is said to be *identified* in a causal model if it is a function of the observed data distribution $p(\mathbf{V})$. Otherwise the parameter is said to be *non-identified*. In all causal models of a DAG $\mathcal{G}$, all interventional distributions $p(\{\mathbf{V} \setminus \mathbf{A}\}(\mathbf{a}))$ are identified by the *g-formula* [11]:

$$p(\{\mathbf{V} \setminus \mathbf{A}\}(\mathbf{a})) = \prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V|\, \mathrm{pa}_{\mathcal{G}}(V))\big|_{\mathbf{A}=\mathbf{a}} \qquad (2)$$

Not all interventional distributions are identified when there are hidden variables present in the causal model. We discuss identification theory in hidden variable DAGs later in this paper.

Edge Interventions

A more general type of intervention in a graphical causal model is the *edge intervention* [15], which maps a set of directed edges in $\mathcal{G}$ to values of their source vertices. Edge interventions have a natural interpretation in cases where a treatment variable has multiple components that a) influence the outcome in different ways, b) occur or do not occur together in observed data, and c) may in principle be intervened on separately. For instance, smoking leads to poor health outcomes due to two components: smoke inhalation and exposure to nicotine. A smoker would be exposed to both of these components, while a non-smoker to neither. However, one might imagine exposing someone selectively only to nicotine but not smoke inhalation (via a nicotine patch), or only smoke inhalation but not nicotine (via smoking plant matter not derived from tobacco leaves). These types of hypothetical experiments correspond precisely to edge interventions, and have been used to conceptualize direct and indirect effects [8, 12], often on the mean difference scale.

Formally, we will write the mapping of a set of edges to values of their source vertices using the following shorthand: $(a_1 W_1)_{\rightarrow}, (a_2 W_2)_{\rightarrow}, \ldots, (a_k W_k)_{\rightarrow}$ to mean that edge $(A_1 W_1)_{\rightarrow}$ is assigned to value a_1, $(A_2 W_2)_{\rightarrow}$ is assigned to value a_2, and so on until $(A_k W_k)_{\rightarrow}$ is assigned to value a_k. Alternatively, we will write $\mathfrak{a}_\alpha$ to mean edges in α are mapped to values in the *multiset* $\mathfrak{a}$ (since multiple edges may share the same source vertex, and be assigned to different values). For a subset $\beta \subseteq \alpha$, and an assignment $\mathfrak{a}_\alpha$ denote $\mathfrak{a}_\beta$ to be a restriction of $\mathfrak{a}_\alpha$ to edges in β.

We will write counterfactual responses to edge interventions as $Y(\mathfrak{a}_\alpha)$ or, for simple cases, as: $Y((aY)_{\rightarrow}, (a'M)_{\rightarrow})$ meaning the response to Y where A is set to value a for the purposes of the edge $(AY)_{\rightarrow}$ and to a' for the purposes of the edge $(AM)_{\rightarrow}$. An edge intervention that sets a set of edges α to values in the multiset $\mathfrak{a}$ is defined via the following generalization of recursive substitution (1):

$$Y(\mathfrak{a}_\alpha) \equiv Y(\mathfrak{a}_{\{(ZY)_{\rightarrow} \in \alpha\}}, \{\mathrm{pa}^{\bar{\alpha}}_{\mathcal{G}}(Y)\}(\mathfrak{a}_\alpha)), \qquad (3)$$

where $\mathrm{pa}^{\bar{\alpha}}_{\mathcal{G}}(Y) \equiv \{W \mid (WY)_{\rightarrow} \notin \alpha\}$. For example, in the DAG in Fig. 1 (a), $Y((a'Y)_{\rightarrow}, (aM)_{\rightarrow})$ is defined as $Y(a', M(a, W), W)$.

For simplicity of presentation, we will restrict attention to edge interventions with the property that if $(AW)_{\rightarrow} \in \alpha$, then for any $V \in \mathrm{ch}_{\mathcal{G}}(A)$, $(AV)_{\rightarrow} \in \alpha$. These types of edge interventions set values for all causal pathways for a set of treatment variables. This is the convention in the majority of existing mediation literature as these interventions are most relevant in practical mediation analysis problems. Specifically, in our HIV example, we are interested in the effect of a drug along all pathways that start with a particular edge, while the effect of the drug via pathways that begin with other edges is kept to a reference level. This assumption may be relaxed, at the price of complicating the theory [15].

Edge interventions are used to define direct and indirect effects. For example, in the model given by the DAG in Fig 1 (a), the direct effect of A on Y is defined as $\mathbb{E}[Y((aY)_{\rightarrow}, (aM)_{\rightarrow})] - \mathbb{E}[Y((a'Y)_{\rightarrow}, (aM)_{\rightarrow})]$ which is equal to $\mathbb{E}[Y(a)] - \mathbb{E}[Y(a', M(a))]$. The indirect effect may be defined similarly as $\mathbb{E}[Y((a'Y)_{\rightarrow}, (aM)_{\rightarrow})] - \mathbb{E}[Y((a'Y)_{\rightarrow}, (a'M)_{\rightarrow})]$, which is equal to $\mathbb{E}[Y(a', M(a))] - \mathbb{E}[Y(a')]$. The direct and indirect effects add up to the ACE.

Note that while direct, indirect, and path-specific effects may be defined directly as nested counterfactuals [8, 13], this notation quickly becomes unreadable for complicated interventions applied at multiple time points. The edge intervention notation may be viewed as a generalization of the do(.) operator notation of Pearl to mediation problems, which avoids having to specify the entire nested counterfactual, and instead directly ties interventions and sets of causal pathways to which these interventions apply (as represented by the first edge shared by all pathways in the set).

Identification of edge interventions in graphical causal models without hidden variables corresponds quite closely with identification of regular (node) interventions, as follows. Let $\mathbf{A}_\alpha \equiv \{A \mid (AB)_{\rightarrow} \in \alpha\}$. Consider an edge intervention given by the mapping $\mathfrak{a}_\alpha$. Then, under the functional model of a DAG $\mathcal{G}$, the joint distribution of counterfactual responses $p(\{\mathbf{V} \setminus \mathbf{A}_\alpha\}(\mathfrak{a}_\alpha))$ is identified via the the following generalization of (2) called the *edge g-formula*:

$$\prod_{V \in \mathbf{V} \setminus \mathbf{A}_\alpha} p(V | \mathfrak{a}_{\{(ZV)_{\rightarrow} \in \alpha\}}, \mathrm{pa}^{\bar{\alpha}}_{\mathcal{G}}(V)). \quad (4)$$

For example, in Fig 1 (a), $p(Y((aY)_{\rightarrow}, (a'M)_{\rightarrow})) = \sum_{W,M} p(Y|a, M, W)p(M|a', W)p(W)$, which is obtained by marginalizing W, M from the edge g-formula.

Edge interventions represent a special case of the more general notion of a *path intervention* [15]. Responses to both of these interventions are used to define *path-specific effects* [8], however responses to edge interventions are precisely those that are always identified under the functional model of a DAG, via (3). Responses to path interventions that cannot be rephrased as responses to edge interventions are not identified even in a DAG model, including the functional model, due to the presence of *recanting witnesses* [1]. For this reason, in this paper we restrict attention only to edge interventions and responses to edge-specific policies.

Responses To Treatment Policies

In personalized medicine settings, counterfactual responses to conditional interventions that set treatment values in response to other variables via a known function are of interest. As an example, assume the graph in Fig. 1 (b) represents an observational study of cancer patients where W_0 represents baseline patient metrics, A_1 is the primary therapy, W_1 is the measured intermediate response to the primary therapy, A_2 is a decision to either continue primary therapy or switch to a secondary therapy in the event of a poor response to A_1, and W_2 is the outcome of interest. In this setting, we might be interested in evaluating policies in the set $\{f_{A_1} : \mathfrak{X}_{W_0} \mapsto \mathfrak{X}_{A_1}, f_{A_2} : \mathfrak{X}_{\{W_0, W_1\}} \mapsto \mathfrak{X}_{A_2}\}$ that map patient characteristics to decisions about therapies A_1 and A_2. We evaluate the efficacy of these policies via the counterfactual variable $W_2(f_{A_1}, f_{A_2})$, representing patient outcomes had treatment decisions been made according to those policies.

These types of variables are defined via a generalization of (1), where instead of setting values of parents in A_1, A_2 to values fixed by the intervention, values of parents in A are instead set according to f_{A_1} and f_{A_2}. In particular, $W_2(f_{A_1}, f_{A_2})$ is defined as

$$W_2[f_{A_2}(W_1[f_{A_1}(W_0), W_0], W_0), W_1[f_{A_1}(W_0), W_0], f_{A_1}(W_0), W_0]. \quad (5)$$

The distribution of this variable is identified under the functional model via the natural generalization of (2) as

$$\sum_{W_0, W_1} p(W_2|W_0, f_{A_1}(W_0), W_1, f_{A_2}(W_0, W_1)) \times p(W_1|W_0, f_{A_1}(W_0))p(W_0). \quad (6)$$

More generally, given a DAG $\mathcal{G}$, a topological ordering $\prec$, and a set $\mathbf{A} \subseteq \mathbf{V}$, for each $A \in \mathbf{A}$, define $\mathbf{W}_A$ to be some subset of predecessors of A according to $\prec$. Then, given a set of functions $\mathbf{f_A}$ of the form $f_A : \mathfrak{X}_{\mathbf{W}_A} \mapsto \mathfrak{X}_A$, define $Y(\mathbf{f_A})$, the counterfactual response $Y \in \mathbf{V}$ to $\mathbf{A}$ being intervened on via $\mathbf{f_A} \equiv \{f_A \mid A \in \mathbf{A}\}$, as

$$Y(\{f_A(\mathbf{W}_A(\mathbf{f_A}))|A \in \mathrm{pa}_{\mathcal{G}}(Y) \cap \mathbf{A}\}, \{\mathrm{pa}_{\mathcal{G}}(Y) \setminus \mathbf{A}\}(\mathbf{f_A})). \quad (7)$$

In a functional model of a DAG $\mathcal{G}$, the effect of $\mathbf{f_A}$ on the set of variables not being intervened upon, $\mathbf{V} \setminus \mathbf{A}$, represented by the distribution $p(\{\mathbf{V} \setminus \mathbf{A})\}(\mathbf{f_A}))$, is identified by the following modification of (2) [16]:

$$\prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V|\{f_A(\mathbf{W}_A)|A \in \mathbf{A} \cap \mathrm{pa}_{\mathcal{G}}(V)\}, \mathrm{pa}_{\mathcal{G}}(V) \setminus \mathbf{A}). \quad (8)$$

3 EDGE-SPECIFIC POLICIES

We now give a general definition of counterfactual responses to edge-specific policies that generalize both responses to edge interventions (where a variable is set to

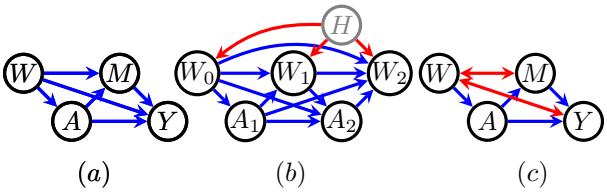

Figure 1: (a) A simple causal DAG, with a treatment A, an outcome Y, a vector W of baseline variables, and a mediator M. (b) A more complex causal DAG with two treatments A_1, A_2, an intermediate outcome W_1, and the final outcome W_2. H is a hidden common cause of the W variables. (c) A graph where $p(Y(a, M(a')))$ is identified, but $p(Y(f_A(W), M(a)))$ is not.

different constants for different outgoing edges) and responses to policies, where a variable is set according to a single known function for all causal pathways at once.

As an example, we can view Fig. 1 (a) as representing a cross-sectional study of HIV patients of the kind described in [6], where W is a set of baseline characteristics, A is one of a set of possible antiretroviral treatments, M is adherence to treatment, and Y is a binary outcome variable signifying viral failure. In this type of study, we may wish to find $f_A(W)$ that maximizes the expected outcome Y had A been set according to $f_A(W)$ for the purposes of the direct effect of A on Y, and A were set to some reference level a for the purposes of the effect of A on M. In other words, we may wish to find $f_A(W)$ to maximize the counterfactual mean $\mathbb{E}[Y(f_A(W), M(a, W), W)]$. This would correspond to finding a treatment policy that maximizes the direct (chemical) effect, if it were possible to keep adherence to a level $M(a)$ as if a reference (easy to adhere to) treatment a were given.

We now give a general definition for responses to such edge-specific policies. Fix a set of directed edges α, and define $\mathbf{A}_\alpha \equiv \{A \mid (AB)_\rightarrow \in \alpha\}$. As before, we assume if $(AW)_\rightarrow \in \alpha$, then for all $V \in \mathrm{ch}_\mathcal{G}(A)$, $(AV)_\rightarrow \in \alpha$. Define $\mathfrak{f}_\alpha \equiv \{f_A^{(AW)_\rightarrow} : \mathfrak{X}_{\mathbf{W}_A} \mapsto \mathfrak{X}_A \mid (AW)_\rightarrow \in \alpha\}$ as the set of policies associated with edges in α. Note that $\mathfrak{f}_\alpha$ may contain multiple policies for a given treatment variable A.

Define $Y(\mathfrak{f}_\alpha)$, the counterfactual response of Y to the set of edge-specific policies $\mathfrak{f}_\alpha$, as the following generalization of (3) and (7):

$$Y(\{f_A^{(AY)_\rightarrow}(\mathbf{W}_A(\mathfrak{f}_\alpha))|(AY)_\rightarrow \in \alpha\}, \{\mathrm{pa}_\mathcal{G}^{\bar{\alpha}}(Y)\}(\mathfrak{f}_\alpha)) \quad (9)$$

In our earlier example, if $\mathfrak{f}_{\{(AY)_\rightarrow,(AM)_\rightarrow\}} \equiv \{f_A^{(AY)_\rightarrow}(W), \tilde{f}_A^{(AM)_\rightarrow}\}$, where $\tilde{f}_A$ assigns A to a constant value a, then $Y(\mathfrak{f}_{\{(AY)_\rightarrow,(AM)_\rightarrow\}}) \equiv Y(f_A(W), M(a, W), W)$.

The joint counterfactual distribution for responses to edge-specific policies, $p(\{V(\mathfrak{f}_\alpha)|V \in \mathbf{V} \setminus \mathbf{A}_\alpha\})$, is identified under the functional model, and generalizes (4) and (6) as follows:

$$\prod_{V \in \mathbf{V}\setminus\mathbf{A}_\alpha} p(V|\{f_A^{(AV)_\rightarrow}(\mathbf{W}_A)|(AV)_\rightarrow \in \alpha\}, \mathrm{pa}_\mathcal{G}^{\bar{\alpha}}(V)). \quad (10)$$

This is a consequence of the fact that (4) holds regardless of how edge interventions are set. In Fig. 1 (a), for example, $p(Y(f_A(W), M(a, W), W)) = \sum_{W,M} p(Y|f_A(W), M, W)p(M|a, W)p(W)$.

4 IDENTIFICATION IN HIDDEN VARIABLE DAG MODELS

In a causal model of a DAG where some variables are hidden, not every causal parameter is a function of the observed data distribution. It is well known, however, that any two hidden variable DAGs which share a special mixed graph called a *latent projection* [9] share identification theory (see [10] for a proof).

Given a DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$, where $\mathbf{V}$ are observed and $\mathbf{H}$ are hidden variables, define a latent projection $\mathcal{G}(\mathbf{V})$ to be an acyclic directed mixed graph (ADMG) with the vertex set $\mathbf{V}$ and $\rightarrow$ and $\leftrightarrow$ edges. An edge $A \rightarrow B$ exists in $\mathcal{G}(\mathbf{V})$ if there is a directed path from A to B in $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ with all intermediate vertices in $\mathbf{H}$. Similarly, an edge $A \leftrightarrow B$ exists in $\mathcal{G}(\mathbf{V})$ if there is a path without consecutive edges $\rightarrow \circ \leftarrow$ from A to B with the first edge on the path of the form $A \leftarrow$ and the last edge on the path of the form $\rightarrow B$, and all intermediate vertices on the path in $\mathbf{H}$. For example, the graph in Fig. 2 (b) is the latent projection of Fig. 2 (a).

We will describe identification results on latent projections directly. General algorithms for identification of interventional distributions were given in [14, 17], for responses to edge interventions in [13], and for policies in [16]. Here we reformulate these results as one line formulas using the fixing operator described in [10]. We do so to explicate the connection between these earlier results, and our new identification algorithm.

Reformulation Of The ID Algorithm

A complete algorithm, called the ID algorithm, for identifying interventional distributions of the form $p(\mathbf{Y}|\mathrm{do}(\mathbf{a}))$, or $p(\mathbf{Y}(\mathbf{a}))$, for $\mathbf{Y} \subseteq \mathbf{V} \setminus \mathbf{A}$ was given in [17] and simplified in [14]. We now illustrate how this algorithm may be further simplified into a one line formula, which can be viewed as a generalization of the g-formula from the fully observed DAG to the hidden variable DAG case. We then show how this formula may

be generalized appropriately to yield identification algorithms for edge interventions, and edge-specific policies in hidden variable causal models, just as g-formula was generalized to these cases in fully observed DAGs.

The version of the ID algorithm in [14], shown in Fig. 1 in the Appendix, proceeds as follows. Lines 2 and 3 reformulate the original query $p(\mathbf{Y}(\mathbf{a}))$ as $\sum_{\mathbf{Y}^* \setminus \mathbf{Y}} p(\mathbf{Y}^*(\mathbf{a}^*))$, where $\mathbf{Y}^*, \mathbf{A}^*$ partition $\text{an}_{\mathcal{G}}(\mathbf{Y})$, and $\mathbf{Y}^* \equiv \text{an}_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}}}(\mathbf{Y})$. In line 4, the distribution $p(\mathbf{Y}^*(\mathbf{a}^*))$ is factorized into terms corresponding to districts $\mathbf{D}$ in the subgraph $\mathcal{G}_{\mathbf{Y}^*}$, with the ID algorithm called recursively on each term. These terms correspond to interventional distributions $p(\mathbf{D} \mid \text{do}(\mathbf{V} \setminus \mathbf{D} = \mathbf{c}_{\mathbf{V}\setminus\mathbf{D}}))$, where $\mathbf{c}_{\mathbf{V}\setminus\mathbf{D}}$ is any set of values of $\mathbf{V} \setminus \mathbf{D}$ consistent with $\mathbf{a}$. In subsequent recursive calls, lines 2, 6 and 7 are iterated for each term until it is identified, or the failure condition is reached. Here line 2 corresponds to marginalizing out irrelevant variables, and lines 6 and 7 correspond to identifying a part of the set of intervened on variables in $\mathbf{V} \setminus \mathbf{D}$ via the g-formula.

Consider Fig. 2 (b), where A represents a binary treatment, Y an outcome of interest, W_0 a vector of baseline confounding factors, and M, W_1 variables mediating the causal effect of A on Y. We are interested in identifying the counterfactual distribution $p(Y(a))$ as a function of the observed data distribution $p(W_0, A, M, W_1, Y)$. Here $\text{an}_{\mathcal{G}}(Y) = \{Y, M, W_1, W_0, A\}$ is partitioned into $\mathbf{Y}^* \equiv \{Y, M, W_1, W_0\}$ and $\mathbf{A}^* \equiv \{A\}$, with $\mathcal{G}_{\mathbf{Y}^*}$ shown in Fig. 2 (c). There are three districts in this graph, $\{W_0, M\}$, $\{W_1\}$, and $\{Y\}$. Thus, the ID algorithm attempts to identify $p(W_0, M|\text{do}(w_1, y, a))$, $p(W_1|\text{do}(w_0, m, y, a))$ and $p(Y|\text{do}(w_0, m, w_1, a))$.

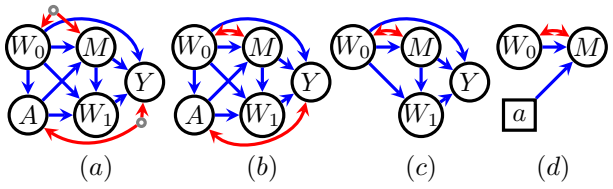

Figure 2: (a) A causal model with a treatment A and outcome Y. (b) A latent projection of the DAG in (a). (c) The graph derived from (b) corresponding to $\mathcal{G}_{\mathbf{Y}^*} = \mathcal{G}_{\{Y, M, W_0, W_1\}}$. (d) A CADMG corresponding to $p(M, W_0|\text{do}(a))$.

As an example, identifying $p(W_0, M|\text{do}(w_1, y, a))$ entails the following steps. First, Y and W_1, as irrelevant variables that do not cause W_0 and M, are marginalized out via line 2, leading to a subproblem where $p(W_0, M|\text{do}(a))$ is identified from $p(W_0, A, M)$ with the subgraph corresponding to this subproblem shown in Fig. 2 (d). In this subproblem, $p(W_0, M|\text{do}(a))$ is identified as $p(M|a, W_0)p(W_0)$ via the g-formula in line 6. The recursion alternates steps that marginalize and apply the g-formula can be unified via a fixing operator applied to graphs and distributions that arise in the intermediate steps of the ID algorithm. We now define these graphs and distributions formally.

CADMGs And Kernels

A *kernel* $q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$ is a mapping from $\mathfrak{X}_{\mathbf{W}}$ to normalized densities over $\mathbf{V}$. Conditioning and marginalization are defined in kernels in the usual way:

$$q_{\mathbf{V}}(\mathbf{A}|\mathbf{W}) \equiv \sum_{\mathbf{V}\setminus\mathbf{A}} q_{\mathbf{V}}(\mathbf{V}|\mathbf{W}); \; q_{\mathbf{V}}(\mathbf{V}\setminus\mathbf{A}|\mathbf{A}\cup\mathbf{W}) \equiv \frac{q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})}{q_{\mathbf{V}}(\mathbf{A}|\mathbf{W})},$$

for $\mathbf{A} \subseteq \mathbf{V}$. A conditional distribution is one type of kernel, but others are possible. The functional $p(M|a, W_0)p(W_0) = p(W_0, M|\text{do}(a))$ in the previous example is a kernel, $q(M, W_0|a)$, that is not in general equal to the conditional distribution $p(M, W_0|a)$.

A conditional ADMG (CADMG) $\mathcal{G}(\mathbf{V}, \mathbf{W})$ is a type of ADMG where nodes are partitioned into two sets. The set $\mathbf{W}$ corresponds to *fixed* constants, and the set $\mathbf{V}$ corresponds to *random variables*. A CADMG has the property that no edges with an arrowhead into an element of $\mathbf{W}$ may exist. Intuitively, a CADMG represents a situation where some variables have already been intervened on. Pearl introduced a similar concept called the 'mutilated graph' in [9]. For example, the graph in Fig. 2 (d) is a CADMG $\mathcal{G}(\{W_0, M\}, \{A\})$ corresponding to the situation where W_0, M are random variables and A is fixed to a constant. Just as a distribution may be associated with a DAG via factorization, so may a kernel be associated with a CADMG in a particular way [10]. For example, the CADMG in Fig. 2 (d) may be associated with $p(W_0, M|\text{do}(a)) = p(M|a, W_0)p(W_0)$. Genealogic definitions, such as $\text{pa}_{\mathcal{G}}(.)$, carry over identically to CADMGs. Districts in a CADMG are defined as subsets of $\mathbf{V}$.

The Fixing Operator And The ID Algorithm

Given a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$, a variable $V \in \mathbf{V}$ is *fixable* if $\text{de}_{\mathcal{G}}(V) \cap \text{dis}_{\mathcal{G}}(V) = \emptyset$. For example, in Fig. 2 (b), M is fixable, while W_0 is not. Intuitively, V is fixable in a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$ if, in a causal graph representing a hypothetical situation $p(\mathbf{V}|\text{do}(\mathbf{w}))$, where variables in $\mathbf{W}$ were already intervened on, $p(\mathbf{V} \setminus \{V\}|\text{do}(\mathbf{w}, v))$ is identified by the application of the g-formula to $p(\mathbf{V}|\text{do}(\mathbf{w}))$. Whenever a variable V is fixable, a fixing operator may be applied to both the CADMG and the kernel to yield a new causal graph and a new kernel representing the situation where V is also intervened on.

Given $V \in \mathbf{V}$ fixable in a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$, the fixing operator $\phi_V(\mathcal{G})$ yields a new CADMG $\widetilde{\mathcal{G}}(\mathbf{V} \setminus \{V\}, \mathbf{W} \cup \{V\})$, where all vertices and edges in $\mathcal{G}(\mathbf{V}, \mathbf{W})$ are kept, *except* V is viewed as fixed, and all edges with arrowheads into V are removed. Given $V \in \mathbf{V}$ fixable in a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$, and a kernel $q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$ associated with $\mathcal{G}$, the fixing operator $\phi_V(q_{\mathbf{V}}; \mathcal{G})$ yields a new kernel $\tilde{q}_{\mathbf{V}\setminus\{V\}}(\mathbf{V} \setminus \{V\}|\mathbf{W} \cup \{V\}) \equiv q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})/q_{\mathbf{V}}(V|\mathbf{W} \cup \mathrm{nd}_{\mathcal{G}}(V))$, where the denominator is defined as above by marginalization and conditioning within the kernel $q_{\mathbf{V}}$. If $\mathrm{ch}_{\mathcal{G}}(V) = \emptyset$, division by $q_{\mathbf{V}}(V|\,\mathrm{nd}_{\mathcal{G}}(V))$ is equivalent to marginalizing V from $q_{\mathbf{V}}$. In this way, the fixing operator unifies applications of the g-formula in lines 6 and 7 of the ID algorithm, and marginalization of irrelevant variables in line 2 of the ID algorithm, and the recursive operation of the ID algorithm can be expressed concisely as repeated invocations of the operator. This allows us to concisely express functionals returned by ID algorithm and its variations, including our new algorithm for identifying responses to edge-specific policies, as one line formulas.

A set $\mathbf{V}^\dagger \subseteq \mathbf{V}$ is said to be *fixable* in a latent projection $\mathcal{G}(\mathbf{V})$ if there is a *valid* sequence $\langle V_1, V_2, \ldots, V_k \rangle$ of variables in $\mathbf{V}^\dagger$ such that V_1 is fixable in $\mathcal{G}$, V_2 is fixable in $\phi_{V_1}(\mathcal{G})$, and so on. If $\mathbf{V}^\dagger$ is fixable, $\mathbf{V} \setminus \mathbf{V}^\dagger$ is called a *reachable* set. If $p(\mathbf{V})$ is a marginal of a distribution $p(\mathbf{V} \cup \mathbf{H})$ Markov relative to a DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$, and $\mathcal{G}(\mathbf{V})$ is a latent projection, then CADMG/kernel pairs obtained from $\mathcal{G}(\mathbf{V})$ and $p(\mathbf{V})$ by any valid sequence in $\mathbf{V}^\dagger$ is the same [10, 17]. As a result, for any fixable set $\mathbf{V}^\dagger$ in $\mathcal{G}$, writing $\phi_{\mathbf{V}^\dagger}(\mathcal{G})$ or $\phi_{\mathbf{V}^\dagger}(q_{\mathbf{V}}; \mathcal{G})$ is well-defined, and means "apply the fixing operator to elements of $\mathbf{V}^\dagger$ in some valid sequence," with the understanding that any such sequence will yield the same result.

The existence of a valid fixing sequence for each district in $\mathcal{G}_{\mathbf{Y}^*}$ implies corresponding terms may be identified via lines 2, 6, and 7 of the ID algorithm, and the overall algorithm can be rephrased as:

$$p(\mathbf{Y}|\mathrm{do}(\mathbf{a})) = \sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} p(\mathbf{D}|\mathrm{do}(\mathbf{V} \setminus \mathbf{D}))|_{\mathbf{A}=\mathbf{a}} \quad (11)$$
$$= \sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} \phi_{\mathbf{V}\setminus\mathbf{D}}(p(\mathbf{V}); \mathcal{G}(\mathbf{V}))|_{\mathbf{A}=\mathbf{a}},$$

which yields the following identifying formula for $p(Y|\mathrm{do}(a))$ in our example in Fig. 2 (a):

$$p(Y(a)) = \sum_{W_0, A, M, W_1} p(W_1|M, A=a, W_0) \times \quad (12)$$
$$p(M|A=a, W_0)p(W_0) \sum_{W_0, A} p(Y|W_1, M, A, W_0)p(W_0, A).$$

We omit the full derivation in the interest of space. See the section on identification of edge-specific policy interventions and the appendix for a complete example. Observe that this equation is a generalized version of Pearl's front-door formula [9].

Whenever $\mathbf{V} \setminus \mathbf{D}$ for every $\mathbf{D}$ is fixable, the formula (11) yields the correct expression for $p(\mathbf{Y}|\mathrm{do}(\mathbf{a}))$ in terms of the observed data. If some $\mathbf{V} \setminus \mathbf{D}$ is not fixable, the algorithm fails, and $p(\mathbf{Y}|\mathrm{do}(\mathbf{a}))$ is not identified. See [10] for a detailed proof.

Edge Interventions

Identification of path-specific effects where each path is associated with one of two possible value sets $\mathbf{a}, \mathbf{a}'$ was given a general characterization in [13] via the *recanting district criterion*. Here, we reformulate this result in terms of the fixing operator in a way that generalizes (11), and applies to the response of any edge intervention, including those that set edges to multiple values rather than two. This result can also be viewed as a generalization of *node consistency* of edge interventions in DAG models, found in [15].

Given $\mathbf{A}_\alpha \equiv \{A \mid (AB)_\rightarrow \in \alpha\}$, and an edge intervention given by the mapping $\mathfrak{a}_\alpha$, define $\mathbf{Y}^* \equiv \mathrm{an}_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}_\alpha}}(\mathbf{Y})$. The joint distribution of the counterfactual response $p(\{\mathbf{V} \setminus \mathbf{A}_\alpha\}(\mathfrak{a}_\alpha))$ is identified if $p(\{\mathbf{V} \setminus \mathbf{A}_\alpha\}(\mathbf{a}))$ is identified via (11), and for every $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})$, for every $A \in \mathbf{A}_\alpha$, $\mathfrak{a}_\alpha$ has the same value assignment for every directed edge out of A into $\mathbf{D}$. Under these assumptions, we have the following result.

Theorem 1 *$p(\mathbf{Y}(\mathfrak{a}_\alpha))$ is identified and equal to*

$$\sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} \phi_{\mathbf{V}\setminus\mathbf{D}}(p(\mathbf{V}); \mathcal{G})\big|_{\mathfrak{a}_{\{(AD)_\rightarrow \in \alpha \mid D \in \mathbf{D}, A \in \mathbf{A}_\alpha\}}} \quad (13)$$

Proof: This follows directly from results in [13] and [10]. Identifying edge interventions entails identifying $\prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} p(\mathbf{D}|\mathrm{do}(\mathbf{a}_\mathbf{D}))$, where $\mathbf{a}_\mathbf{D}$ is an assignment for $\mathrm{pa}^{\mathrm{s}}_{\mathcal{G}}(\mathbf{D})$, and $\mathbf{a}_\mathbf{D}$ possibly assigns different values to elements of $\mathbf{A}$ with respect to different districts. The fact that this identification algorithm can be rephrased as (13) follows directly by Theorem 60 in [10]. □

Consider again the example in Fig. 2 (a). Now assume we set $A = a$ for the edge $(AM)_\rightarrow$ and $A = a'$ for the edge $(AW_1)_\rightarrow$. The identifying functional for $p(Y((aW_1)_\rightarrow, (a'M)_\rightarrow))$ has the same form as (12), but some terms are evaluated at $A = a$, and some at $A = a'$:

$$\sum_{W_0, A, M, W_1} p(W_1|M, A=a, W_0) \quad (14)$$
$$p(M|A=a', W_0)p(W_0) \sum_{W_0, A} p(Y|W_0, A, M, W_1)p(W_0, A)$$

Policy Interventions (Dynamic Treatment Regimes)

A general algorithm for identification of responses to a set of policies $\mathbf{f_A}$ was given in [16]. We again reformulate this algorithm in terms of the fixing operator. Define a graph $\mathcal{G}_{\mathbf{f_A}}$ to be a graph obtained from $\mathcal{G}$ by removing all edges into $\mathbf{A}$, and adding for any $A \in \mathbf{A}$, directed edges from $\mathbf{W}_A$ to A. By definition of $\mathbf{W}_A$, $\mathcal{G}_{\mathbf{f_A}}$ is guaranteed to be acyclic. Define $\mathbf{Y}^* \equiv \mathrm{an}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{Y}) \setminus \mathbf{A}$. Assume $p(\mathbf{Y}^*(\mathbf{a}))$ is identified in $\mathcal{G}$. Then, under the above assumptions, we have the following result.

Theorem 2 *$p(\mathbf{Y}(\mathbf{f_A}))$ is identified in $\mathcal{G}$. Moreover, the identification formula is*

$$\sum_{(\mathbf{Y}^* \cup \mathbf{A}) \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} \phi_{\mathbf{V} \setminus \mathbf{D}}(p(\mathbf{V}); \mathcal{G})\big|_{\tilde{\mathbf{a}}_{\mathrm{pa}^s_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}}} \quad (15)$$

where $\tilde{\mathbf{a}}_{\mathrm{pa}^s_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}}$ is defined as

$$\begin{cases} \{A = f_A(\mathbf{W}_A) \mid A \in \mathrm{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}\} & \mathrm{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A} \neq \emptyset \\ \emptyset & \textit{otherwise} \end{cases}$$

Proof: This follows from the fact that identification of $p(\mathbf{Y}(\mathbf{f_A}))$ can be rephrased as identification of $p(\mathbf{Y}^*(\mathbf{a}))$, with values $\mathbf{a}$ set according to $\{\mathbf{W}_A | A \in \mathbf{A}\}$, where all $\mathbf{W}_A$ in the set are subsets of $\mathbf{Y}^*$. Identification of $p(\mathbf{Y}^*(\mathbf{a}))$ may be rephrased as (15) follows by Theorem 60 in [10]. □

The outer sum over $\mathbf{A}$ in (15) is vacuous if $\mathbf{f_A}$ is a set of deterministic policies. To illustrate (15), in our example in Fig. 2 (b), $p(Y(A = f_A(W_0)))$ is identified as

$$\sum_{W_0, A, M, W_1} p(W_1|M, A{=}f(W_0), W_0) \quad (16)$$
$$p(M|A{=}f(W_0), W_0) p(W_0) \sum_{W_0, A} p(Y|W_1, M, A, W_0) p(W_0, A)$$

Identification Of Edge-Specific Policies

Having reformulated existing identification results on responses to policies (15) and responses to edge interventions arising in mediation analysis (13) in terms of the fixing operator, we generalize these results for identification of responses to edge-specific policies.

Given $\mathbf{A}_\alpha \equiv \{A | (AB)_\rightarrow \in \alpha\}$, and a set of edge-specific policies given by the set of mappings $\mathfrak{f}_\alpha$, define the graph $\mathcal{G}_{\mathfrak{f}_\alpha}$ to be one where all edges with arrowheads into $\mathbf{A}_\alpha$ are removed, and directed edges from any vertex in $\mathbf{W}_A$ to $A \in \mathbf{A}_\alpha$ added. Fix a set $\mathbf{Y}$ of outcomes of interest, and define $\mathbf{Y}^*$ equal $\mathrm{an}_{\mathcal{G}_{\mathfrak{f}_\alpha}}(\mathbf{Y}) \setminus \mathbf{A}_\alpha$. We have the following result.

Theorem 3 *$p(\mathbf{Y}(\mathfrak{f}_\alpha))$ is identified if $p(\mathbf{Y}^*(\mathbf{a}))$ is identified, and for every $\mathbf{D} \in \mathcal{D}((\mathcal{G}_{\mathfrak{f}_\alpha})_{\mathbf{Y}^*})$, $\mathfrak{f}_\alpha$ yields the same policy assignment for every edge from $A \in \mathbf{A}_\alpha$ to $\mathbf{D}$. Moreover, the identifying formula is*

$$\sum_{(\mathbf{Y}^* \cup \mathbf{A}_\alpha) \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} \phi_{\mathbf{V} \setminus \mathbf{D}}(p(\mathbf{V}); \mathcal{G})|_{\tilde{\mathbf{a}}_{\mathrm{pa}^s_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}_\alpha}} \quad (17)$$

where $\tilde{\mathbf{a}}_{\mathrm{pa}^s_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}_\alpha}$ is defined to be $\{A = f_A(\mathbf{W}_A) \in \mathfrak{f}_\alpha \mid A \in \mathrm{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}_\alpha\}$, if $\mathrm{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}_\alpha \neq \emptyset$, and is defined to be the $\emptyset$ otherwise.

Proof: This is a straightforward generalization of the proofs of Theorems 1 and 2. □

Responses to edge-specific policies are identified in strictly fewer cases compared to responses to edge interventions. This is because $\mathbf{Y}^*$ is a larger set in the former case. As an example, consider the graph in Fig. 1 (c), where we are interested either in the counterfactual $p(Y(a, M(a')))$, used to define pure direct effects, or the counterfactual $p(Y(f_A(W), M(a')))$.

For the former counterfactual, we have $\mathbf{Y}^* = \{Y, M\}$, and $p(Y(a, M(a')))$ equal to

$$\sum_m \left(\frac{\sum_w p(Y, m|a, w) p(w)}{\sum_w p(m \mid a, w) p(w)} \right) \sum_w p(m \mid a', w) p(w)$$

We omit the detailed derivation in the interest of space. For the latter counterfactual, however, the set $\mathbf{Y}^* = \{Y, M, W\}$ forms a single district in $\mathcal{G}_{\mathbf{Y}^*}$, and the edge-specific policy set $\mathfrak{f}_{\{(AM)_\rightarrow, (AY)_\rightarrow\}}$ sets edges from A to this district to different policies. As a result, Theorem 3 is insufficient to conclude identification.

Generalizations of the example in Fig. 1 (b) are the most relevant in practice, as their causal structure corresponds to longitudinal observational studies, of the kind considered in [11], and many other papers. However, we illustrate complications that may arise in identifiability of responses to edge-specific policies with our running example in Fig. 2 (b), where we are interested in the response of Y to edge-specific policies $\mathfrak{f}_{\{(AM)_\rightarrow, (AW_1)_\rightarrow\}} = \{f_A^{(AM)_\rightarrow}(W_0), f_A^{(AW_1)_\rightarrow}(W_0)\}$. Theorem 3 yields the following identifying formula:

$$\sum_{W_0, A, M, W_1} \Big[\big[p(W_1|M, A = f_A^{(AM)_\rightarrow}(W_0), W_0) \big] \quad (18)$$
$$\times \big[p(M|A = f_A^{(AW_1)_\rightarrow}(W_0), W_0) p(W_0) \big]$$
$$\times \big[\sum_{W_0, A} p(Y|W_1, M, A, W_0) p(W_0, A) \big] \Big].$$

Note that (18) generalizes both (14), which sets A to different constants in different terms, and (16), which sets A to the output of a function that depends on W_0. We give a detailed derivation of this functional in the appendix.

5 ON COMPLETENESS

An identification algorithm for a class of parameters is said to be *complete* relative to a class of causal models if, whenever the algorithm fails to identify a parameter within a model class, the parameter is in fact not identified within that class.

The ID algorithm is known to be complete for the class of interventional distributions in the class of functional models [5, 14]. We restate this result here, and give a sequence of increasingly general completeness results for the identification algorithms described so far. Completeness results on policies and edge-specific policies are new. For completeness results pertaining to policies, we assume a completely unrestricted class of policies. If the set of policies of interest, $\mathbf{f_A}$ or $\mathfrak{f}_\alpha$ is restricted, or alternatively if the causal model has parametric restrictions, completeness results we present may no longer hold.

Theorem 4 *Given disjoint subsets $\mathbf{Y}, \mathbf{A}$ of $\mathbf{V}$ in an ADMG $\mathcal{G}$, define $\mathbf{Y}^* \equiv \mathrm{an}_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}}}(\mathbf{Y})$. Then $p(\mathbf{Y}(\mathbf{a}))$ is not identified if there exists $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})$ that is not a reachable set in $\mathcal{G}$.*

Corollary 1 *The algorithm for identification of $p(\mathbf{Y}(\mathbf{a}))$, as phrased in (11), is complete.*

Theorem 5 *Given $\mathbf{A}_\alpha \equiv \{A \mid (AB)_\rightarrow \in \alpha\}$, and an edge intervention given by the mapping $\mathfrak{a}_\alpha$, define $\mathbf{Y}^* \equiv \mathrm{an}_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}_\alpha}}(\mathbf{Y})$. The joint distribution of the counterfactual response $p(\{\mathbf{V} \setminus \mathbf{A}_\alpha\}(\mathfrak{a}_\alpha))$ is not identified if $p(\{\mathbf{V} \setminus \mathbf{A}_\alpha\}(\mathbf{a}))$ is not identified, or there exists $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})$ and $A \in \mathbf{A}_\alpha$, such that $\mathfrak{a}_\alpha$ has the different value assignments for a pair of directed edges out of A into $\mathbf{D}$.*

Corollary 2 *The algorithm for identification of $p(\mathbf{Y}(\mathfrak{a}_\alpha))$, as phrased in (13), is complete.*

Theorem 6 *Define $\mathcal{G}_{\mathbf{f_A}}$ to be a graph obtained from $\mathcal{G}$ by removing all edges into $\mathbf{A}$, and adding for any $A \in \mathbf{A}$, directed edges from $\mathbf{W}_A$ to A. Define $\mathbf{Y}^* \equiv \mathrm{an}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{Y}) \setminus \mathbf{A}$. Then if $p(\mathbf{Y}^*(\mathbf{a}))$ is not identified in $\mathcal{G}$, $p(\mathbf{Y}(\mathbf{f_A}))$ is not identified in $\mathcal{G}$ if $\mathbf{f_A}$ is the unrestricted class of policies.*

Corollary 3 *The algorithm for identification of $p(\mathbf{Y}(\mathbf{f_A}))$, as phrased in (15), is complete for unrestricted policies.*

Theorem 7 *Define the graph $\mathcal{G}_{\mathfrak{f}_\alpha}$ to be one where all edges with arrowheads into $\mathbf{A}_\alpha$ are removed, and directed edges from any vertex in $\mathbf{W}_A$ to $A \in \mathbf{A}_\alpha$ added. Fix a set $\mathbf{Y}$ of outcomes of interest, and define $\mathbf{Y}^*$ equal $\mathrm{an}_{\mathcal{G}_{\mathfrak{f}_\alpha}}(\mathbf{Y}) \setminus \mathbf{A}_\alpha$. Then if $p(\mathbf{Y}^*(\mathbf{a}))$ is not identified, or there exists $\mathbf{D} \in \mathcal{D}((\mathcal{G}_{\mathfrak{f}_\alpha})_{\mathbf{Y}^*})$, such that $\mathfrak{f}_\alpha$ yields different policy assignments for two edges from $A \in \mathbf{A}_\alpha$ to $\mathbf{D}$, $p(\mathbf{Y}(\mathfrak{f}_\alpha))$ is not identified.*

Corollary 4 *The algorithm for identification of $p(\mathbf{Y}(\mathfrak{f}_\alpha))$, as phrased in (17), is complete for unrestricted policies.*

Detailed proofs of these results are in the Appendix. Corollaries are immediate consequences of the preceding Theorems.

6 CONCLUSION

In this paper, we defined counterfactual responses to policies that set treatment values in such a way that they affect outcomes with respect to certain causal pathways only. Such counterfactuals arise when we wish to personalize only some portion of the causal effect of a treatment, while keeping other portions set to some reference values. An example might be optimizing the chemical effect of a drug, while keeping drug adherence to a reference value.

We gave a general algorithm for identifying these responses from data, which generalizes similar algorithms due to [16, 13] for dynamic treatment regimes, and edge-specific effects, respectively. Further, we showed that given an unrestricted class of policies the algorithm is complete. As a corollary, this established that the identification algorithm for dynamic treatment regimes in [16] is complete for unrestricted policies.

Given a fixed set of policies associated with a set of causal pathways, and assuming (17) yields a functional containing only conditional densities, as is the case in the functional (18), the counterfactual mean under those policies $\mathbb{E}[Y(\mathfrak{f}_\alpha)]$ may be estimated using the maximum likelihood plug-in estimator. Such an estimator can be viewed as a generalization of the parametric g-formula [11] to edge-specific policies. More general estimation strategies, and approaches to learning the optimal set of policies are the subject of our companion paper [7].

Acknowledgments

This research was supported in part by the NIH grants R01 AI104459-01A1 and R01 AI127271-01A1. We thank the anonymous reviewers for their insightful comments that greatly improved this manuscript.

References

[1] C. Avin, I. Shpitser, and J. Pearl. 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, 2005.

[2] D. P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Publishing, 1996.

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

[4] M. A. Hernan, E. Lanoy, D. Costagliola, and J. M. Robins. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic and Clinical Pharmacology and Toxicology*, 98:237–242, 2006.

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

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

[7] R. Nabi and I. Shpitser. Estimation of personalized effects associated with causal pathways. In *Proceedings of the Thirty Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

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

[9] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

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

[11] J. M. Robins. 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, 1986.

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

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

[14] I. Shpitser and J. Pearl. 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, 2006.

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

[16] J. Tian. 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, 2008. AUAI Press.

[17] J. Tian and J. Pearl. 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, 2002.