Modeling Latent Selection with Structural Causal Models

Leihao Chen
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Amsterdam, the Netherlands

Onno Zoeter
Booking.com
The Netherlands

Joris M. Mooij
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Amsterdam, the Netherlands

Abstract

Selection bias is ubiquitous in real-world data, and can lead to misleading results if not dealt with properly. We introduce a conditioning operation on Structural Causal Models (SCMs) to model latent selection from a causal perspective. We show that the conditioning operation transforms an SCM with the presence of an explicit latent selection mechanism into an SCM without such selection mechanism, which partially encodes the causal semantics of the selected subpopulation according to the original SCM. Furthermore, we show that this conditioning operation preserves the simplicity, acyclicity, and linearity of SCMs, and commutes with marginalization. Thanks to these properties, combined with marginalization and intervention, the conditioning operation offers a valuable tool for conducting causal reasoning tasks within causal models where latent details have been abstracted away. We demonstrate by example how classical results of causal inference can be generalized to include selection bias and how the conditioning operation helps with modeling of real-world problems.

1 Introduction

Selection bias is prevalent in numerous real-world problems, and naive analyses can result in counterintuitive paradoxes and misleading conclusions (Berkson, 1946; Zhao et al., 2021; Fryer Jr, 2019). While there are various methods to address selection bias (see, e.g., Smith (2020) and the references therein), a causal perspective can aid in understanding its behavior structurally (Hernán et al., 2004; Bareinboim and Pearl, 2012).

One approach to modeling selection bias involves using a Structural Causal Model (SCM, see Pearl (2009)) that explicitly describes the selection mechanism, as demonstrated in the 's-recoverability' work by Bareinboim and Pearl (2012). However, this necessitates a detailed knowledge of the selection mechanism, which is often unavailable. In this context, we propose an alternative method for addressing *latent selection bias* that allows one to abstract away modeling details irrelevant to the causal inference task of interest.

To illustrate, we discuss a toy example, demonstrating how to obtain correct results without assuming any specific details about the latent selection mechanism.

Example 1 (Car mechanic) *Cars start successfully if their battery is charged and their start engine is operational. Introduce binary endogenous variables B_0 ("battery"), E_0 ("start engine") and S_0 ("car starts") measured at time t_0 and variables B_1, E_1 and S_1 with similar meaning for the same car but measured at time t_1 with $t_1 > t_0$. Assume that the following SCM, whose graph is depicted in Figure 1, is a causal model for the population of all cars:*[1]

$$M : \begin{cases} U_B \sim \mathrm{Ber}(1-\delta), U_E \sim \mathrm{Ber}(1-\epsilon), \\ B_0 = U_B, E_0 = U_E, S_0 = B_0 \wedge E_0, \\ B_1 = B_0, E_1 = E_0, S_1 = B_1 \wedge E_1. \end{cases}$$

where U_B and U_E are latent exogenous independent Bernoulli-distributed random variables with parameters $1-\delta$ and $1-\epsilon$. Assume that states of the battery and start engine do not change from time t_0 to t_1.

A car mechanic can use this model M to predict the effects of interventions on cars. For example, the probability that charging the battery will make non-starting cars start is $\mathrm{P}_M(S_1 = 1 \mid \mathrm{do}(B_1 = 1), S_0 = 0) =$

[1] For simplicity, we assume no other possible causes of a non-starting car (like the fuel tank being empty).

$\frac{\delta(1-\epsilon)}{\delta+(1-\delta)\epsilon}$, *and the probability that replacing the start engine will make non-starting cars start is* $\mathrm{P}_M(S_1 = 1 \mid \mathrm{do}(E_1 = 1), S_0 = 0) = \frac{(1-\delta)\epsilon}{\epsilon+\delta(1-\epsilon)}$.

Now suppose that the car mechanic is ignorant of the model M, but wants to take a data-driven approach to repairing cars. The goal is to use a large number of observational and interventional data that she collected from her workshop to estimate an SCM $\tilde{M}$ that allows her to predict the effects of interventions (charging the battery, replacing the start engine, etc.) on non-starting cars. The SCM $\tilde{M}$ that she estimates, whose graph is depicted in Figure 1, is given by

$$\tilde{M} : \begin{cases} (U_B, U_E) \sim \mathrm{P}_\theta(U_B, U_E) \\ B_1 = U_B, E_1 = U_E, S_1 = B_1 \wedge E_1 \end{cases}$$

with $\mathrm{P}_\theta(U_B, U_E) = \frac{\delta\epsilon}{\delta+(1-\delta)\epsilon}\delta_{\{U_B=U_E=0\}} + \frac{\delta(1-\epsilon)}{\delta+(1-\delta)\epsilon}\delta_{\{U_B=0,U_E=1\}} + \frac{(1-\delta)\epsilon}{\delta+(1-\delta)\epsilon}\delta_{\{U_B=1,U_E=0\}}$. *Performing calculations in $\tilde{M}$ gives the two target interventional quantities* $\mathrm{P}_{\tilde{M}}(S_1 = 1 \mid \mathrm{do}(B_1 = 1)) = \frac{\delta(1-\epsilon)}{\delta+(1-\delta)\epsilon}$ *and* $\mathrm{P}_{\tilde{M}}(S_1 = 1 \mid \mathrm{do}(E_1 = 1)) = \frac{(1-\delta)\epsilon}{\epsilon+\delta(1-\epsilon)}$. *Besides this, $\tilde{M}$ reproduces the observational distribution for non-starting cars:* $\mathrm{P}_{\tilde{M}}(B_1, E_1, S_1) = \mathrm{P}_M(B_1, E_1, S_1 \mid S_0 = 0)$. *So, the car mechanic (who might not even be aware of the latent selection mechanism $S_0 = 0$) can still use an SCM as an accurate causal model to predict the effects of interventions (on the subpopulation of cars that are of her concern).*

Note that we could also have obtained the model $\tilde{M}$ directly from M, by (i) replacing $\mathrm{P}_M(U_B, U_E)$ by $\mathrm{P}_M(U_B, U_E \mid S_0 = 0)$, and (ii) marginalizing out B_0, E_0 and S_0 (by substituting the structural equations for B_0, E_0 and S_0 in the remaining structural equations and then removing these variables from the model). This allows us to effectively abstract away irrelevant latent modeling details: (i) the latent variables B_0, E_0 and S_0, (ii) their causal mechanisms, and (iii) the selection step on $S_0 = 0$. In Section 3, we formally define this operation of *conditioning on an event* for SCMs in more generality.

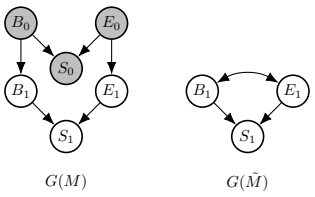

Figure 1: The causal graphs of the SCMs M and $\tilde{M}$ in Example 1. Conditioning M on $S_0 = 0$ yields $\tilde{M}$. The gray nodes are latent.

Our Contributions We define a conditioning operation on *SCMs* to model latent selection. The observational probability distribution of the conditioned SCM is the same as the one of the original SCM conditional on the selection event. The conditioned SCM also preserves interventional and counterfactual semantics w.r.t. the non-ancestors of the selection nodes, and preserves important model properties (acyclicity/linearity/simplicity). For the graphical representation of SCMs, we generalize the semantics of bidirected edges in directed mixed graphs so that these do not only indicate latent common confounding but also indicate dependencies stemming from latent selection. Our construction allows to easily extend many existing results for SCMs (e.g., do-calculus, adjustment criteria, identification results, ect.) to allow for conditioning and latent selection.

Related Work Bareinboim and Tian (2015) dealt with the so-called 's-recoverability' problem of selection bias. They made an explicit modeling assumption about the selection variables via causal graphs and explored graphical criteria under which one can recover causal quantities in the total population from the data of the subpopulation. There exist probabilistic graphical model classes equipped with marginalization and conditioning operations, e.g., ancestral graphs (Richardson and Spirtes, 2002) and d- or σ-connection graphs (Hyttinen et al., 2014; Forré and Mooij, 2018) whose causal interpretation is not completely clear. For example, Zhang (2008) gave a causal interpretation to MAGs ruling out selection bias, but no follow-up work deals with selection bias, as far as we know.

2 Preliminaries

In this section, we recall some basics of SCMs and introduce notations we use in the current paper. To save space, we put some definitions in supplement. We follow the formal setup of Bongers et al. (2021).

Definition 2 (Structural Causal Model) *A **Structural Causal Model (SCM)** is a tuple $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ such that*

- *V, W are disjoint finite sets of labels for the **endogenous variables** and the **exogenous random variables**, respectively;*
- *the **state space** $\mathcal{X} = \prod_{i \in V \cup W} \mathcal{X}_i$ is a product of standard measurable spaces $\mathcal{X}_i$;*
- *the **exogenous distribution** P is a probability distribution on $\mathcal{X}_W$ that factorizes as a product $\mathrm{P} = \bigotimes_{w \in W} \mathrm{P}(X_w)$ of probability distributions $\mathrm{P}(X_w)$ on $\mathcal{X}_w$;*
- *the **causal mechanism** is specified by the measurable mapping $f : \mathcal{X} \to \mathcal{X}_V$.*

Definition 3 (Hard intervention) *Given an SCM M, an intervention target $T \subseteq V$ and an intervention value $x_T \in \mathcal{X}_T$, we define the* ***intervened SCM***

$$M_{\mathrm{do}(X_T = x_T)} := (V, W, \mathcal{X}, \mathrm{P}, (f_{V\setminus T}, x_T)).$$

This replaces the targeted endogenous variables by specified values. In this work, we do not assume that all the endogenous variables in an SCM can be intervened on, which deviates from the standard modeling assumption. One can define other types of interventions like soft or probabilistic ones, and the results in the following also hold replacing hard interventions by other types of interventions.

Besides interventional semantics, one can also describe counterfactual semantics of an SCM by performing interventions in its twin SCM (see supplement).

Given an SCM M, one can define its causal graph $G(M)$ and augmented causal graph $G^a(M)$ to give an intuitive and compact graphical representation of the causal model (see supplement). One can read off useful causal information purely from the causal graphs without knowing the details of the underlying SCMs.

Notation 4 *In all the causal graphs, we use gray nodes to represent latent variables. We assume that latent variables are non-intervenable. Dashed nodes represent observable but non-intervenable variables, and solid nodes represent observable and intervenable variables. Exogenous variables are assumed latent.*

Definition 5 (Solution function of an SCM) *Given an SCM M, we call a measurable mapping $g^S : \mathcal{X}_{V\setminus S} \times \mathcal{X}_W \to \mathcal{X}_S$ a* ***solution function of*** *M* ***w.r.t.*** *$S \subseteq V$ if for $\mathrm{P}(X_W)$-a.a. $x_W \in \mathcal{X}_W$ and for all $x_{V\setminus S} \in \mathcal{X}_{V\setminus S}$, one has that $g^S(x_{V\setminus S}, x_W)$ satisfies the structural equations for S, i.e.,*

$$g^S(x_{V\setminus S}, x_W) = f_S(x_{V\setminus S}, g^S(x_{V\setminus S}, x_W), x_W).$$

When $S = V$, we denote g^V by g, and call g a ***solution function of*** *M.*

Definition 6 (Unique solvability) *An SCM M is called* ***uniquely solvable w.r.t.*** *$S \subseteq V$ if it has a solution function w.r.t. S that is* ***essentially unique*** *in the sense that if g^S and $\tilde{g}^S$ both satisfy the structural equations for S, then for $\mathrm{P}(X_W)$-a.a. $x_W \in \mathcal{X}_W$ and for all $x_{V\setminus S} \in \mathcal{X}_{V\setminus S}$, one has $g^S(x_{V\setminus S}, x_W) = \tilde{g}^S(x_{V\setminus S}, x_W)$. If M has an essentially unique solution function w.r.t. V, then it is called* ***uniquely solvable****.*

Note that a subset S does not inherit unique solvability from unique solvability of any of its supersets in general (Bongers et al. (2021, Appendix B.2)).

Definition 7 (Simple SCM) *An SCM M is called a* ***simple SCM*** *if it is uniquely solvable w.r.t. each subset $S \subseteq V$.*

Simple SCMs form a class of SCMs that preserves most convenient properties of acyclic SCMs but allows for weak cycles (acyclic SCMs are simple). We focus on simple SCMs in this work so that we can avoid many mathematical technicalities interfering with intuition. We use $\mathrm{P}_M(X_V, X_W)$ to denote the unique probability distribution of (X_V, X_W) induced by a simple SCM M.

For a simple SCM, we can plug the solution function of one component into other parts of the model so that we can get a simple SCM that "marginalizes" it out while preserving causal semantics (Bongers et al., 2021).

Definition 8 (Marginalization) *Let M be a simple SCM and $L \subseteq V$. Then we call $M_{\setminus L} = (V \setminus L, W, \mathcal{X}_{V\setminus L} \times \mathcal{X}_W, \mathrm{P}, \tilde{f})$ with*

$$\tilde{f}(x_{V\setminus L}, x_W) = f_{V\setminus L}(x_{V\setminus L}, g^L(x_{V\setminus L}, x_W), x_W)$$

a ***marginalization*** *of M over $V \setminus L$.*

3 Conditioning Operation on SCMs

In Section 3.1, we define the conditioning operation on simple SCMs. We shall derive some properties of it in Section 3.2, and the proofs can be found in supplement. In Section 3.3, we make some important caveats on how to interpret the conditioned SCMs when modeling. In the whole section, we assume:

Assumption 9 *$M = (V, W, \mathcal{X}, \mathrm{P}, f)$ is a simple SCM such that $\mathrm{P}_M(X_S \in \mathcal{S}) > 0$ for some $S \subseteq V$ and measurable subset $\mathcal{S} \subseteq \mathcal{X}_S$.*

We write $O := V \setminus S$. We use $\mathrm{P}_M(X_O \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S}) := \mathrm{P}_{M_{\mathrm{do}(X_T = x_T)}}(X_O \mid X_S \in \mathcal{S})$ to represent the probability distribution of X_O when first intervening on $X_T = x_T$ and second conditioning on $X_S \in \mathcal{S}$.[2]

3.1 Definition of Conditioning Operation

Suppose that we condition on the event $\{X_S \in \mathcal{S}\}$. Then roughly speaking, the conditioning operation can be divided into three steps:

1. merging all the exogenous random variables that are ancestors of the selection variables;
2. updating the exogenous probability distribution to the posterior given the observation $X_S \in \mathcal{S}$;

[2] "First" and "second" here refer to the order of applying the operations on the SCM.

3. marginalizing out the selection variables.

We give the formal definition of the conditioned SCMs specializing to the class of simple SCMs for simplicity. See Figure 2 for an intuitive graphical representation.[3]

Definition 10 (Conditioned SCM) *Assume Assumption 9. Write* $S^c := (V \cup W) \setminus S$ *and* $B := \mathrm{Anc}_{G^a(M)}(S)$. *Let* $g^S : \mathcal{X}_W \times \mathcal{X}_O \to \mathcal{X}_S$ *be the essentially unique solution function of* M *w.r.t.* S. *We define the* ***conditioned SCM*** $M_{|X_S \in \mathcal{S}} := \left(\hat{V}, \hat{W}, \hat{\mathcal{X}}, \hat{\mathrm{P}}, \hat{f}\right)$ *by:*

- $\hat{V} := V \setminus S$;
- $\hat{W} := (W \setminus B) \dot\cup \{\star_W\}$ *with* $\star_W := B \cap W$;
- $\hat{\mathcal{X}} := \mathcal{X}_O \times \hat{\mathcal{X}}_{\hat{W}} := \mathcal{X}_O \times \left(\mathcal{X}_{W\setminus B} \times \mathcal{X}_{\star_W}\right)$, *where* $\mathcal{X}_{\star_W} := \mathcal{X}_{W\cap B}$;
- $\hat{\mathrm{P}} := \mathrm{P}(X_{W\setminus B}) \otimes \mathrm{P}(X_{\star_W})$, *where* $\mathrm{P}(X_{\star_W}) := \mathrm{P}_M(X_{W\cap B} \mid X_S \in \mathcal{S})$;
- $\hat{f}(x_{\hat{V}}, x_{\hat{W}}) := f_O(x_O, g^S(x_O, x_{W\setminus B}, x_{\star_W}), x_{W\setminus B}, x_{\star_W})$.

It is easy to check that $M_{|X_S\in\mathcal{S}}$ is indeed an SCM.

Remark 11 *(1) In Definition 10,* $M_{|X_S\in\mathcal{S}}$ *actually depends on the choice of* g^S, *but different versions are equivalent (in the sense of Definition 32). Here we abuse terminology and call* $M_{|X_S\in\mathcal{S}}$ *"the conditioned SCM" of* M *given* $X_S \in \mathcal{S}$ *rather than "a conditioned SCM", and work with equivalence classes of SCMs. Note that if* M *and* $\tilde{M}$ *are equivalent, then* $M_{|X_S\in\mathcal{S}}$ *is equivalent to* $\tilde{M}_{|X_S\in\mathcal{S}}$.

(2) The reasons of assuming $\mathrm{P}(X_S \in \mathcal{S}) > 0$ *are twofold. First, in the real world, we never observe data from a null event, and therefore it is reasonable not to model such cases. Second, conditioning on null events will introduce mathematical technicalities.*

(3) Since marginalization preserves simplicity, $M_{|X_S\in\mathcal{S}}$ *is simple (see also Proposition 14).*

(4) Since M *is simple, we only need to update the exogenous distribution of ancestors of the selection variables, which is not the case for nonsimple SCMs. It is possible to generalize all the results in this work to more general class of SCMs.*

[3] In the graphical representations of the conditioning operation such as Figure 2 and Figure 3, we assume no causal effects canceled out because of marginalization or changing the underlying population.

Notation 12 *We often denote* $M_{|X_S\in\mathcal{S}}$ *by* $M_{|\mathcal{S}}$ *if it is clear from the context that* $\mathcal{S}$ *is a measurable subset in which the variable* X_S *takes values.*

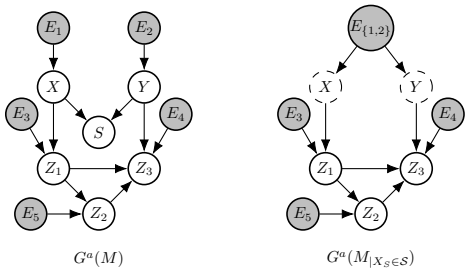

Figure 2: Graphical representation of conditioning on $X_S \in \mathcal{S}$. First merge the exogenous ancestors of S, i.e., E_1 and E_2, to get a merged node $E_{\{1,2\}}$. Then update the exogenous probability distribution $\mathrm{P}(X_{E_1}, X_{E_2})$ to the posterior $\mathrm{P}_M(X_{E_1}, X_{E_2} \mid X_S \in \mathcal{S})$. Finally, marginalize out the node S. X and Y are dashed, since they have become non-intervenable (see Section 3.3).

If $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$ with measurable subsets $\mathcal{S}_i \subseteq \mathcal{X}_{S_i}$ for $i = 1, \ldots, n$, then we could also define $M_{|\mathcal{S}} := \left((M_{|\mathcal{S}_1})\ldots\right)_{|\mathcal{S}_n}$ by applying the conditioning operation iteratively to the components of the Cartesian product instead. This can give a more fine-grained model as shown in Figure 3. The problem is that in general applying the iterative conditioning in different orders may result in non-equivalent models with different graphs. Luckily, iterative conditioning will generate counterfactually equivalent SCMs, as we will show in the next subsection.[4]

Note that the above conditioning operation is defined on simple SCMs. In practice, people often use causal graphs to communicate causal knowledge, without referring to the underlying SCMs. To support this, we give a purely graphical conditioning operation defined on directed mixed graphs (DMGs). An example is given in supplement. As we will show in the next subsection, the purely graphical conditioning operation interacts well with the SCM conditioning operation.

Definition 13 (Conditioned DMG) *Let* $G = (V, E, H)$ *be a DMG consisting of nodes* V, *directed edges* E *and bidirected edges* H. *For* $S \subseteq V$, *we define the conditioned DMG as*

$$G_{|S} = \left(V_{|S}, E_{|S}, H_{|S}\right),$$

where

- $V_{|S} := V \setminus S$ *with* $\mathrm{Anc}_G(S) \setminus S$ *dashed;*

[4] Note that counterfactually equivalent SCMs may not be equivalent or possess the same graphs. See Bongers et al. (2021).

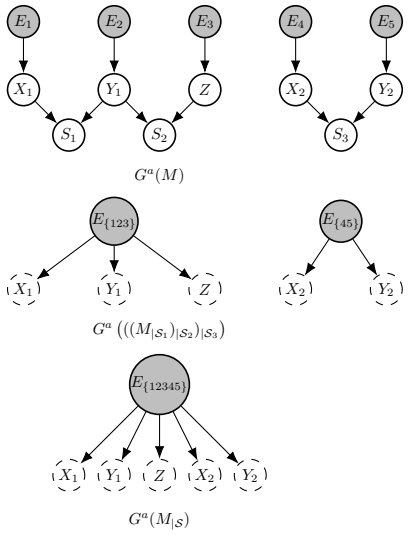

Figure 3: Graphical representation of the multiple-node conditioning operation. If we condition on the event $(S_1, S_2, S_3) \in \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3$ via iterative single-node conditioning, then we get a finer conditioned SCM $((M_{|\mathcal{S}_1})_{|\mathcal{S}_2})_{|\mathcal{S}_3}$ in contrast to a coarser model $M_{|\mathcal{S}}$ obtained by a single (joint) conditioning operation.

- $E_{|S}$ *consists of all* $v \longrightarrow u$ *with* $v, u \in V \setminus S$ *and* $v \neq u$ *for which there exists a directed walk in* G*:* $v \longrightarrow s_1 \longrightarrow \cdots \longrightarrow s_n \longrightarrow u$*, where all intermediate nodes* $s_1, \ldots, s_n \in S$ *(if any);*
- $H_{|S}$ *consists of all bidirected edges* $v \longleftrightarrow u$ *with* $v, u \in V \setminus S$ *and* $v \neq u$*, for which there exists a bifurcation in* G*:* $v \longleftarrow s_1 \longleftarrow \cdots \longleftarrow s_{k-1} \longleftrightarrow s_k \longrightarrow \cdots \longrightarrow s_n \longrightarrow u$ *with all intermediate nodes* $s_1, \ldots, s_n \in S$ *(if any), or for which* $v, u \in \mathrm{Anc}_G(S) \cup \mathrm{Sib}_G(\mathrm{Anc}_G(S))$.[5]

3.2 Properties of Conditioning Operation

The conditioning operation preserves simplicity, linearity, and acyclicity of SCMs.

Proposition 14 (Simplicity/Acyclicity/Linearity) *If* M *is a simple (resp. acyclic) SCM with conditioned SCM* $M_{|X_S \in \mathcal{S}}$*, then the conditioned SCM* $M_{|X_S \in \mathcal{S}}$ *is simple (resp. acyclic). If* M *is also linear, then so is* $M_{|X_S \in \mathcal{S}}$.

The following proposition states that the purely graphical conditioning operation is compatible with the SCM conditioning operation.

Proposition 15 (Conditioned SCM & DMG) *Let* M *be a simple SCM with conditioned SCM* $M_{|X_S \in \mathcal{S}}$*. Then* $G(M_{|X_S \in \mathcal{S}})$ *is a subgraph of* $G(M)_{|S}$.

[5] $\mathrm{Sib}_G(v) := \{w \in G \mid v \longleftrightarrow w \in G\}$.

Remark 16 *Note that* $G(M_{|X_S \in \mathcal{S}})$ *can be a strict subgraph of* $G(M)_{|S}$.

The following lemma states that the conditioning commutes with interventions on the non-ancestors of the conditioning variables.

Lemma 17 (Conditioning & intervention) *Assume Assumption 9. Then we have* $\left(M_{\mathrm{do}(X_T = x_T)}\right)_{|X_S \in \mathcal{S}} = \left(M_{|X_S \in \mathcal{S}}\right)_{\mathrm{do}(X_T = x_T)}$ *for any* $T \subseteq O \setminus \mathrm{Anc}_{G^a(M)}(S)$ *and* $x_T \in \mathcal{X}_T$.

Remark 18 *Since* M *is simple and* $T \subseteq O \setminus \mathrm{Anc}_{G^a(M)}(S)$*, the probability* $\mathrm{P}_{M_{\mathrm{do}}(X_T = x_T)}(X_S \in \mathcal{S}) = \mathrm{P}_M(X_S \in \mathcal{S})$ *is well defined and strictly larger than zero.*

The following theorem shows that the conditioned SCM indeed encodes the causal semantics according to the original SCM for the selected subpopulation and for interventions targeting the non-ancestors of S.

Theorem 19 (Preserving causal semantics) *Assume Assumption 9. Then we have*

(1) $\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_O) = \mathrm{P}_M(X_O \mid X_S \in \mathcal{S})$*;*

(2) for any $T \subseteq V \setminus \mathrm{Anc}_{G^a(M)}(S)$ *and* $x_T \in \mathcal{X}_T$*,*

$$\begin{aligned}&\mathrm{P}_{M_{|X_S \in \mathcal{S}}}\left(X_{O \setminus T} \mid \mathrm{do}(X_T = x_T)\right)\\ &= \mathrm{P}_M\left(X_{O \setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S}\right);\end{aligned}$$

(3) for any $T_1 \subseteq V \setminus \mathrm{Anc}_{G^a(M)}(S)$ *and* $x_{T_1} \in \mathcal{X}_{T_1}$*, and any* $T_2 \subseteq (V \setminus \mathrm{Anc}_{G^a(M)}(S))'$ *and* $x_{T_2} \in \mathcal{X}_{T_2}$*,*

$$\begin{aligned}&\mathrm{P}_{\left(M_{|X_S \in \mathcal{S}}\right)^{\mathrm{twin}}}(X_{(O \cup O') \setminus (T_1 \cup T_2)} \mid \\ &\qquad\qquad \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2}))\\ &= \mathrm{P}_{M^{\mathrm{twin}}}(X_{(O \cup O') \setminus (T_1 \cup T_2)} \mid \\ &\qquad\qquad \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2}), X_S \in \mathcal{S}).\end{aligned}$$

The following corollary implies that different orderings of iterative conditioning operations give rise to counterfactually equivalent SCMs.

Corollary 20 (Iterative conditioning) *Assume Assumption 9 with* $S = S_1 \cup S_2$ *and* $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$ *with* $\mathcal{S}_1 \subseteq \mathcal{X}_{S_1}$ *and* $\mathcal{S}_2 \subseteq \mathcal{X}_{S_2}$ *both measurable. Then* $(M_{|\mathcal{S}_1})_{|\mathcal{S}_2}$*,* $(M_{|\mathcal{S}_2})_{|\mathcal{S}_1}$*, and* $M_{|\mathcal{S}_1 \times \mathcal{S}_2}$ *are counterfactually equivalent w.r.t.* $V \setminus \mathrm{Anc}_{G^a(M)}(S_1 \cup S_2)$.[6]

The restriction that $T \cap \mathrm{Anc}_{G^a(M)}(S) = \emptyset$ in Theorem 19 cannot be relaxed, as the following example shows.[7]

[6] See Definition 33 or Bongers et al. (2021, Definition 4.5) for the definition of counterfactual equivalence.

[7] This was also recently observed by Mathur et al. (2023)

Example 21 (Conditioning & intervention) *Consider a linear SCM*

$$M: \begin{cases} T = E_T, \\ X = \alpha T + E_X, \\ Y = X + \beta T + E_Y. \end{cases}$$

The graph $G(M)$ of M is shown in Figure 4.

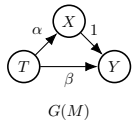

Figure 4: Causal graph of linear SCM in Example 21.

In M, if we first condition on $X = x$ and second intervene on T, then we have

$$\begin{aligned} &\mathbb{E}_{(M_{|X=x})_{\mathrm{do}(T=1)}}[Y] - \mathbb{E}_{(M_{|X=x})_{\mathrm{do}(T=0)}}[Y] \\ &= \mathbb{E}[(X + \beta + E_Y) - (X + E_Y) \mid X = x] \\ &= \beta. \end{aligned}$$

On the contrary, if we first intervene on T and second condition on $X = x$, then we have

$$\begin{aligned} &\mathbb{E}_{(M_{\mathrm{do}(T=1)})_{|X=x}}[Y] - \mathbb{E}_{(M_{\mathrm{do}(T=0)})_{|X=x}}[Y] \\ &= \mathbb{E}_M\left[\alpha + E_X + \beta + E_Y \mid X = x\right] \\ &\quad - \mathbb{E}_M\left[E_X + E_Y \mid X = x\right] \\ &= \alpha + \beta. \end{aligned}$$

In general conditioning and intervention do not commute.

The following results show that conditioning and marginalization commute.

Proposition 22 (Conditioning & marginalization) *Assume Assumption 9 and let $L \subseteq V \setminus S$. Then we have $(M_{\setminus L})_{|\mathcal{S}} = (M_{|\mathcal{S}})_{\setminus L}$.*

3.3 Some Important Caveats on Modeling Interpretation

In the above two subsections, we presented the conditioning operation as a purely mathematical operation and derived some mathematical properties of it. In this subsection, we shall make some remarks on how to interpret the conditioned SCMs appropriately to avoid confusion in modeling applications.

The subtleties are about intervening on ancestors of selection nodes. In this case, conditioning and interventions are not commutative as we showed before. Therefore, one should be careful about the order of these two operations. On the one hand, if we first intervene and second condition on descendants of intervened variables, then the selected subpopulation will also change according to the intervention. On the other hand, first conditioning and second intervening on ancestors of selection nodes has a "counterfactual flavor". Suppose that an SCM M with three variables T ("treatment"), Y ("outcome") and S ("selection") has causal graph $T \dashrightarrow Y \dashrightarrow S$. Intuitively, "first-conditioning-second-intervening" indicates that we first observe the results of the treatment and select units with specific values (say $S = s$) and fix this subpopulation. After that, we go back to then perform an intervention (say $\mathrm{do}(T = t)$) on this *fixed selected subpopulation* instead of the total population. Mathematically, we have

$$\begin{aligned} &\mathrm{P}_{\left((M_{|S=s})_{\mathrm{do}(T=t)}\right)}(Y) \\ &= \mathrm{P}_{M_{|S=s}}(Y \mid \mathrm{do}(T = t)) \\ &= \mathrm{P}_M(Y_t \mid S = s) \\ &= \mathrm{P}_{M^{\mathrm{twin}}}(Y' \mid \mathrm{do}(T' = t), S = s) \\ &\neq \mathrm{P}(Y \mid \mathrm{do}(T = t), S = s), \quad \text{(in general)} \\ &= \mathrm{P}_{\left((M_{\mathrm{do}(T=t))_{|S=s}}\right)}(Y) \end{aligned}$$

where we used the language of potential outcomes. In Pearl's terminology, this mixes different rungs: a rung-two query in the conditioned SCM is equivalent to a rung-three query in the original SCM.

As far as we know, there are two possible ways to use the conditioning operation for modeling without introducing confusion:

- before (or after) performing the conditioning operation, marginalizing out all the ancestors of the selection nodes, so that one can no longer intervene on the ancestors of the selection nodes;
- specifying in the conditioned SCM and its graph which variables are ancestors of the selection nodes in the original SCM, and marking them as non-intervenable (e.g., making them dashed).[8]

Remark 23 *When the selection variables do not have any intervenable ancestors (e.g., all the ancestors of the selection nodes are latent), one can safely apply the conditioning operation without any extra steps.*

4 Some Applications

By the properties of the conditioning operation, all the classical results for SCMs, such as identification results (do-calculus, the back-door criterion), can be applied to conditioned SCMs immediately. Combining with marginalization, the conditioning operation

[8] This means that we obtain a graph with mixed interpretation in the sense that some part of the graph is causal and some part is non-causal (purely probabilistic).

also provides a way for understanding a DMG as a causal graph that compactly encodes causal assumptions, where latent details of both latent confounding and latent selection have been abstracted away.

For illustration purposes, we briefly discuss four examples in this section. They form a cohesive sequence navigating us from high-level philosophical implications of the conditioning operation ("generalized Reichenbach's principle"), to the versatility of applications of classical identification results to conditioned SCMs (Back-door theorem, instrumental variables), and finally low-level concrete practical application of conditioned SCMs in modeling real-world problems (COVID example).

Example 24 (Reichenbach's principle) *Reichenbach's Principle of Common Cause (Reichenbach, 1956) is often stated in this way: if two variables are dependent, then one must cause the other or the variables must have a common cause (or any combination of these three possibilities). It is essential to note that this conclusion holds only when latent selection bias is ruled out, an assumption that is typically left implicit.*

With our conditioning operation, we can generalize it in the following way. Assume that M is a simple SCM that has two observed endogenous variables X and Y. By the Markov property (Bongers et al., 2021, Theorem 6.3), if X and Y are dependent, then $X \dashrightarrow Y$, $X \dashleftarrow Y$, or $X \dashleftarrow\dashrightarrow Y$ (or any combination of these three possibilities) are in the graph $G(M)$. There exist infinitely many SCMs M^i, $i \in I$ with an infinite index set I, such that $(M^i_{\setminus L_i})_{|\mathcal{S}_i} = M$ where L_i is a set of latent variables of M^i and $X_{S_i} \in \mathcal{S}_i$ is the latent selection in M^i. Hence, it implies that if two variables are dependent, then one causes the other, or the variables have a common cause or be subject to latent selection (or any combination of these four possibilities).

Example 25 (Back-door theorem) *Let M^1 and M^2 be two SCMs with three variables T ("treatment"), X ("covariates"), and Y ("outcome") whose causal graphs are shown in Figure 5. Under positivity assumptions, Pearl's Back-Door Theorem (Pearl, 2009) gives, for $i = 1, 2$, the identification result:*

$$\begin{aligned} &\mathrm{P}_{M^i}(Y \mid \mathrm{do}(T = t)) \\ &= \int \mathrm{P}_{M^i}(Y \mid X = x, T = t)\mathrm{P}_{M^i}(X \in \mathrm{d}x). \end{aligned} \tag{1}$$

Thanks to marginalization and the conditioning operation, we can see M^1 and M^2 as abstractions of other SCMs, i.e., $M^i = (\tilde{M}^i_{\setminus L_i})_{|\mathcal{S}_i}$, for SCMs $\tilde{M}^i$, latent variables $L^i = \{L^i_1, \ldots, L^i_n\}$, and latent selection variables $S^i = \{S^i_1, \ldots, S^i_m\}$ taking values in measurable

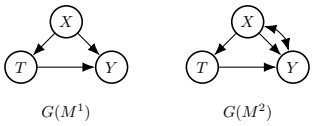

Figure 5: Causal graphs of SCMs M^i in Example 25.

sets $\mathcal{S}^i$ with $i = 1, 2$. We present $\tilde{M}^i_{(j)}$ for $j = 1, 2$ as four examples of the infinitely many possibilities for $\tilde{M}^i$ in Figure 6.

With the help of Theorem 19, we can write (1) *as*

$$\begin{aligned} \mathrm{P}_{\tilde{M}^i}(Y \mid \mathrm{do}(T = t), X_{S^i} \in \mathcal{S}^i) = \int \mathrm{P}_{\tilde{M}^i}(Y \mid X = x, \\ T = t, X_{S^i} \in \mathcal{S}^i)\mathrm{P}_{\tilde{M}^i}(X \in \mathrm{d}x \mid X_{S^i} \in \mathcal{S}^i). \end{aligned}$$

Thus, the Back-door theorem can be applied directly to the conditioned SCM, which is useful if the specific latent structure of the SCM is not precisely known. One can generalize other identifcation results similarly.

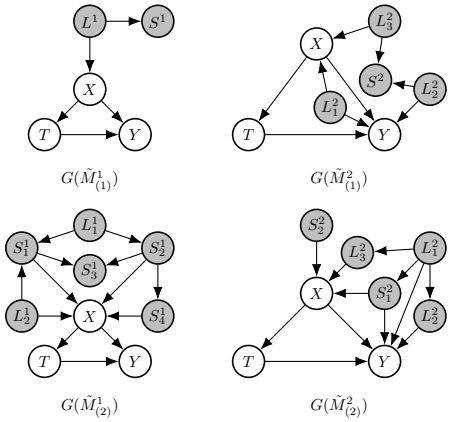

Figure 6: Some possible causal graphs of SCMs $\tilde{M}^i$ in Example 25.

Example 26 (Instrumental variables) *In some situations, one cannot get point identification results but can only derive bounds for target causal effects, e.g., Pearl's instrumental inequality (Pearl, 2009, Chapter 8). It was derived for SCMs with graph $G(M)$ shown in Figure 7. Similar to Example 25, we can generalize the inequality to any SCM $\tilde{M}$ such that for latent variables L and latent selection on $X_S \in \mathcal{S}$, $G\left(\left(\tilde{M}_{\setminus L}\right)_{|\mathcal{S}}\right)$ is of the form shown in Figure 7.*

If we further assume that $Y = \beta X + f(U)$ in M, then when $\mathrm{Cov}(X, Y) \neq 0$ the parameter β is identifiable from the conditional distribution $\mathrm{P}_{\tilde{M}}(T, X, Y \mid X_S \in \mathcal{S})$ and given by $\frac{\mathrm{Cov}_{\tilde{M}}(T,Y|X_S\in\mathcal{S})}{\mathrm{Cov}_{\tilde{M}}(X,Y|X_S\in\mathcal{S})} =: \tilde{\beta}$. That is because $\frac{\mathrm{Cov}_M(T,Y)}{\mathrm{Cov}_M(X,Y)} = \beta = \tilde{\beta}$. Therefore, we have generalized

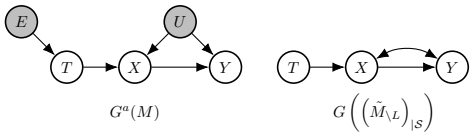

Figure 7: Graphs for the instrumental variables model.

the identification result to including a certain form of selection bias.

Example 27 (Causal modeling) *As concise depictions of causal assumptions, causal graphs guide the selection of appropriate statistical methods for addressing queries. For example, with the aim of understanding the effect of treatment strategies of different countries on the COVID-19 case fatality rate, von Kügelgen et al. (2021) analyzed data from the initial outbreaks in 2020 in China and Italy and assumed a causal graph G shown in Figure 8. For the persons infected with COVID-19, the data record age (A), fatality rate (F), and country of residence (C) at the time of infection. The authors draw a directed edge from C to A to explain the dependency between C and A observed in the data. However, this assumption implies that if we conduct a randomized trail to assign people to different countries, then the resulting age distribution will differ depending on the assigned country, which does not appear to be a reasonable assumption.*

*In fact, we can draw a bidirected edge $C \dashleftarrow\dashrightarrow A$ as shown in $\tilde{G}$ to explain the statistical association between C and A, which could represent different latent selection mechanisms or confounding between C and A. First, the age distribution may differ between two countries already before the outbreak of the virus (latent selection on 'person was alive ($S' = 1$) in early 2020', as in G^1). Second, since only **infected** patients were registered and both country and age may influence the risk of getting infected, selection on infection status ($S = 1$) can also lead to $C \dashleftarrow\dashrightarrow A$ (as in G^2). The combinations of both selection mechanisms (as in G^3 or G^4) also lead to $C \dashleftarrow\dashrightarrow A$. With the conditioning operation, we do not need to list (potentially infinitely many) all the possible **complete** causal graphs including **all** relevant latent variables that model the selection mechanism, and we only need to consider DMGs on these three observed variables, which is a much smaller (finite) model space.*

Thanks to properties of the conditioning operation, we can answer causal queries like "what would be the effect on fatality of changing from China to Italy", i.e., to compute the total causal effect $\mathrm{TCE}(Y; c' \to c) := \mathbb{E}[F \mid \mathrm{do}(C = c)] - \mathbb{E}[F \mid \mathrm{do}(C = c')]$*, via the abstract (conditioned) model $\tilde{G}$ (e.g., via adjusting on age) without fully knowing all the latent details. Note that the results based on G and $\tilde{G}$ are clearly different (see supplement for details).*

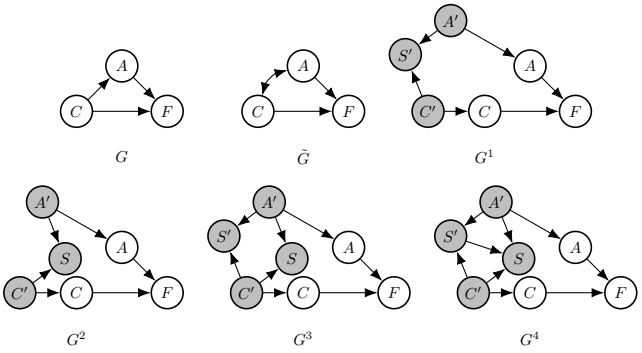

Figure 8: Causal graphs for COVID-19 data. Note that after applying the conditioning operation to selection variables and marginalizing out remaining latent variables, we reduce G^i to $\tilde{G}$ for $i = 1, 2, 3, 4$.

5 Conclusions

While marginalization plays a role of abstracting away unnecessary *unconditioned* latent details of causal models, we need another operation in case of latent selection mechanisms. We gave a formal definition of a conditioning operation on SCMs to take care of latent selection. The conditioning operation preserves a large part of the causal information, preserves important model classes and interacts well with other operations on SCMs. We generalized the interpretation of bidirected edges in directed mixed graphs to represent both latent common causes and selection on a latent event. Combined with marginalization and intervention, the conditioning operation provides a powerful tool for causal model abstraction and helps with many causal inference tasks such as prediction of interventions, identification and model selection.

Acknowledgements

This work is supported by Booking.com. We thank Philip Boeken and Stephan Bongers for discussions.

References

Bareinboim, E. and Pearl, J. (2012). Controlling selection bias in causal inference. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 100–108, La Palma, Canary Islands. PMLR.

Bareinboim, E. and Tian, J. (2015). Recovering causal effects from selection bias. In *Proceedings*

of the AAAI Conference on Artificial Intelligence, volume 29, page 2410–2416.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53.

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. (2021). Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915.

Forré, P. and Mooij, J. M. (2018). Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI-18)*, pages 269–278.

Fryer Jr, R. G. (2019). An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, 127(3):1210–1261.

Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615–625.

Hyttinen, A., Eberhardt, F., and Järvisalo, M. (2014). Constraint-based causal discovery: Conflict resolution with answer set programming. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI-14)*, page 340–349.

Mathur, M. B., Shpitser, I., and VanderWeele, T. (2023). A common-cause principle for eliminating selection bias in causal estimands through covariate adjustment. OSF Preprints ths4e, Center for Open Science.

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

Reichenbach, H. (1956). *The Direction of Time*. Dover Publications, Mineola, N.Y.

Richardson, T. S. and Spirtes, P. (2002). Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030.

Smith, L. H. (2020). Selection mechanisms and their consequences: understanding and addressing selection bias. *Current Epidemiology Reports*, 7:179–189.

von Kügelgen, J., Gresele, L., and Schölkopf, B. (2021). Simpson's paradox in COVID-19 case fatality rates: a mediation analysis of age-related causal effects. *IEEE Transactions on Artificial Intelligence*, 2(1):18–27.

Zhang, J. (2008). Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474.

Zhao, Q., Ju, N., Bacallado, S., and Shah, R. D. (2021). BETS: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic. *The Annals of Applied Statistics*, 15(1):363–390.

The supplementary material contains some definitions and detailed proofs of the results that are missing in the main paper.

Supplement A. More SCM Preliminaries

To be as self-contained as possible, we include the definitions of twin SCM and (augmented) causal graphs of SCMs. We follow the formal definitions of Bongers et al. (2021).

Definition 28 (Twin SCM) *(Bongers et al., 2021, Definition 2.17) Let $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ be an SCM. The twinning operation maps M to the* ***twin structural causal model (twin SCM)***

$$M^{\text{twin}} := \left(V \cup V', W, \mathcal{X}_V \times \mathcal{X}_{V'} \times \mathcal{X}_W, \mathrm{P}, \tilde{f}\right),$$

where $V' = \{v' : v \in V\}$ is a disjoint copy of V and the causal mechanism $\tilde{f} : \mathcal{X}_V \times \mathcal{X}_{V'} \times \mathcal{X}_W \to \mathcal{X}_V \times \mathcal{X}_{V'}$ is the measurable mapping given by $\tilde{f}\left(x_V, x_{V'}, x_W\right) = \left(f(x_V, x_W), f\left(x_{V'}, x_W\right)\right)$.

Definition 29 (Parent) *(Bongers et al., 2021, Definition 2.6) Let $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ be an SCM. We call $k \in V \cup W$ a* ***parent*** *of $v \in V$ if and only if there does not exist a measurable mapping $\tilde{f}_v : \mathcal{X}_{V \setminus k} \times \mathcal{X}_{W \setminus k} \to \mathcal{X}_v$ such that for $\mathrm{P}(X_W)$-almost every $x_W \in \mathcal{X}_W$, for all $x_V \in \mathcal{X}_V$,*

$$x_v = f_v(x_V, x_W) \quad \Longleftrightarrow \quad x_v = \tilde{f}_v\left(x_{V \setminus k}, x_{W \setminus k}\right).$$

Definition 30 (Graph and augmented graph) *(Bongers et al., 2021, Definition 2.7) Let $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ be an SCM. We define:*

(1) the augmented graph $G^a(M)$ as the directed graph with nodes $V \cup W$ and directed edges $u \to v$ if and only if $u \in V \cup W$ is a parent of $v \in V$;

(2) the graph $G(M)$ as the directed mixed graph with nodes V, directed edges $u \dashrightarrow v$ if and only if $u \in V$ is a parent of $v \in V$ and bidirected edges $u \dashleftarrow\dashrightarrow v$ if and only if there exists a $w \in W$ that is a parent of both $u \in V$ and $v \in V$.

Example 31 *Consider the SCM*

$$M : \begin{cases} U \sim \mathrm{Ber}(1-\xi), U_B \sim \mathrm{Ber}(1-\delta), U_E \sim \mathrm{Ber}(1-\varepsilon), \\ B_0 = U, E_0 = U, S_0 = B_0 \wedge E_0, \\ B_1 = B_0 \wedge U_B, E_1 = E_0 \wedge U_E, S_1 = B_1 \wedge E_1. \end{cases}$$

Then we have the (augmented) causal graphs of M shown in Figure 9.

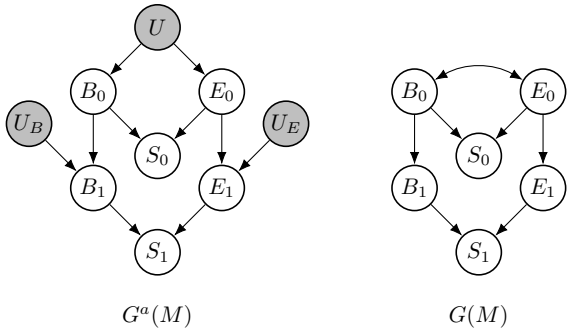

Figure 9: The (augmented) causal graphs of the SCM M in Example 31.

Definition 32 (Equivalence) *(Bongers et al., 2021, Definition 2.5) An SCM $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ is* ***equivalent*** *to an SCM $\tilde{M} = (V, W, \mathcal{X}, \mathrm{P}, \tilde{f})$ if for all $v \in V$, for P-a.a. $x_W \in \mathcal{X}_W$ and for all $x_V \in \mathcal{X}_V$,*

$$x_v = f_v(x_V, x_W) \quad \Longleftrightarrow \quad x_v = \tilde{f}_v(x_V, x_W).$$

Definition 33 (Counterfactual equivalence) *(Bongers et al., 2021, Definition 4.5) An SCM $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ is **counterfactually equivalent** to an SCM $\tilde{M} = (\tilde{V}, \tilde{W}, \tilde{\mathcal{X}}, \tilde{\mathrm{P}}, \tilde{f})$ w.r.t. $O \subseteq V \cap \tilde{V}$ if for any $T_1 \subseteq O$ and $x_{T_1} \in \mathcal{X}_{T_1}$, and any $T_2 \subseteq O'$ and $x_{T_2} \in \mathcal{X}_{T_2}$,*

$$\begin{aligned}&\mathrm{P}_{M^{\mathrm{twin}}}(X_{(O\cup O')\setminus(T_1\cup T_2)} \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2}))\\ &= \mathrm{P}_{\tilde{M}^{\mathrm{twin}}}(X_{(O\cup O')\setminus(T_1\cup T_2)} \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2})).\end{aligned}$$

Supplement B. Proofs

B.1 Proof of Proposition 14

Proposition 14 (Simplicity/Acyclicity/Linearity) *If M is a simple (resp. acyclic) SCM with conditioned SCM $M_{|X_S\in\mathcal{S}}$, then the conditioned SCM $M_{|X_S\in\mathcal{S}}$ is simple (resp. acyclic). If M is also linear, then so is $M_{|X_S\in\mathcal{S}}$.*

Proof We first show that the conditioning operation preserves simplicity of SCMs. First note that marginalization preserves simplicity (Bongers et al., 2021, Proposition 8.2). Also note that both merging exogenous random variables and changing the exogenous probability distribution preserve simplicity. Hence, the conditioning operation preserves simplicity.

We give the proof of the fact that the conditioning operation preserves acyclicity of SCMs. First note that merging exogenous random variables and updating the exogenous probabilistic distribution preserve acyclicity. Then since marginalization preserves acyclicity (Bongers et al., 2021, Proposition 5.11), we get that the conditioning operation preserves acyclicity.

We now show that the conditioning operation preserves linearity of SCMs. Merging exogenous random variables and changing the exogenous probability distribution preserve linearity. Marginalization also preserves linearity (Bongers et al., 2021, Proposition C.5).

B.2 Proof of Proposition 15

Proposition 15 (Conditioned SCM & DMG) *Let M be a simple SCM with conditioned SCM $M_{|X_S\in\mathcal{S}}$. Then $G(M_{|X_S\in\mathcal{S}})$ is a subgraph of $G(M)_{|S}$.*

Proof This is easily seen from the definition and Bongers et al. (2021, Proposition 5.11).

B.3 Proof of Lemma 17

Lemma 17 (Conditioning & intervention) *Assume Assumption 9. Then we have $\left(M_{\mathrm{do}(X_T=x_T)}\right)_{|X_S\in\mathcal{S}} = \left(M_{|X_S\in\mathcal{S}}\right)_{\mathrm{do}(X_T=x_T)}$ for any $T \subseteq O \setminus \mathrm{Anc}_{G^a(M)}(S)$ and $x_T \in \mathcal{X}_T$.*

Proof In the proof, we set $B := \mathrm{Anc}_{G^a(M)}(S)$ and $O := V \setminus S$. We check the definition one by one. For $\left(M_{|\mathcal{S}}\right)_{\mathrm{do}(X_T=x_T)} := (\hat{V}, \hat{W}, \hat{\mathcal{X}}, \hat{\mathrm{P}}, \hat{f})$, we have:

- $\hat{V} = V \setminus S$;
- $\hat{W} = (W \setminus B) \dot{\cup} \{\star_W\}$ with $\star_W = B \cap W$;
- $\hat{\mathcal{X}} = \mathcal{X}_{S^c\setminus(B\cap W)} \times \mathcal{X}_{\star_W}$;
- $\hat{\mathrm{P}} = \hat{\mathrm{P}}\left(X_{W\setminus B}\right) \otimes \hat{\mathrm{P}}(X_{\star_W}) = \mathrm{P}\left(X_{W\setminus B}\right) \otimes \mathrm{P}_M(X_{W\cap B} \mid X_S \in \mathcal{S})$;
- $\hat{f}\left(x_{\hat{V}}, x_{\hat{W}}\right) = \left(f_{O\setminus T}\left(x_O, g^S\left(x_O, x_{W\setminus B}, x_{\star_W}\right), x_{W\setminus B}, x_{\star_W}\right), x_T\right)$.

We write $\tilde{B} := \mathrm{Anc}_{G^a(M_{\mathrm{do}(X_T=x_T)})}(S)$. Note that since $T \cap B = \emptyset$, it follows that $\tilde{B} = B$. Since $T \cap B = T \cap \tilde{B} = \emptyset$, we have $\mathrm{P}_M(X_B) = \mathrm{P}_{M_{\mathrm{do}(X_T=x_T)}}(X_{\tilde{B}})$. Hence, we can conclude that

$$\mathrm{P}_M(X_{W\cap B} \mid X_S \in \mathcal{S}) = \mathrm{P}_{M_{\mathrm{do}(X_T=x_T)}}(X_{W\cap\tilde{B}} \mid X_S \in \mathcal{S}).$$

Combining all the above ingredients, we have for $\left(M_{\mathrm{do}(X_T=x_T)}\right)_{|\mathcal{S}} := (\hat{V}, \hat{W}, \hat{\mathcal{X}}, \hat{\mathrm{P}}, \hat{f})$:

- $\hat{\hat{V}} = V \setminus S$;
- $\hat{\hat{W}} = \left(W \setminus \tilde{B}\right) \dot{\cup} \{\star_W\} = (W \setminus B) \dot{\cup} \{\star_W\}$ with $\star_W = \tilde{B} \cap W = B \cap W$;
- $\hat{\hat{\mathcal{X}}} = \mathcal{X}_{S^c \setminus (\tilde{B} \cap W)} \times \mathcal{X}_{\star_W} = \mathcal{X}_{S^c \setminus (B \cap W)} \times \mathcal{X}_{\star_W}$;
- $\hat{\hat{\mathrm{P}}} = \mathrm{P}(X_{W \setminus \tilde{B}}) \otimes \hat{\mathrm{P}}(X_{\star_W}) = \mathrm{P}(X_{W \setminus \tilde{B}}) \otimes \mathrm{P}_{M_{\mathrm{do}(X_T = x_T)}}(X_{W \cap \tilde{B}} \mid X_S \in \mathcal{S}) = \mathrm{P}(X_{W \setminus B}) \otimes \mathrm{P}_M(X_{W \cap B} \mid X_S \in \mathcal{S})$;
- For the causal mechanism, we have

$$\begin{aligned} \hat{\hat{f}}\left(x_{\hat{\hat{V}}}, x_{\hat{\hat{W}}}\right) &= \tilde{f}_O\left(x_O, \tilde{g}^S\left(x_O, x_{W \setminus \tilde{B}}, x_{\star_W}\right), x_{W \setminus \tilde{B}}, x_{\star_W}\right) \\ &= \left(f_{O \setminus T}\left(x_O, \tilde{g}^S\left(x_O, x_{W \setminus B}, x_{\star_W}\right), x_{W \setminus B}, x_{\star_W}\right), x_T\right), \end{aligned}$$

where $\tilde{f}$ is the causal mechanism of $M_{\mathrm{do}(X_T = x_T)}$ and $\tilde{g}^S$ is the (essentially unique) solution function of $M_{\mathrm{do}(X_T = x_T)}$ w.r.t. S. Note that $g^S = \tilde{g}^S$ as $T \cap B = \emptyset$. Overall, it is then easy to see that $\left(M_{\mathrm{do}(X_T = x_T)}\right)_{|\mathcal{S}} = \left(M_{|\mathcal{S}}\right)_{\mathrm{do}(X_T = x_T)}$.

B.4 Proof of Theorem 19

Theorem 19 (Preserving causal semantics) *Assume Assumption 9. Then we have*

(1) $\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_O) = \mathrm{P}_M(X_O \mid X_S \in \mathcal{S})$;

(2) for any $T \subseteq V \setminus \mathrm{Anc}_{G^a(M)}(S)$ *and* $x_T \in \mathcal{X}_T$,

$$\begin{aligned} &\mathrm{P}_{M_{|X_S \in \mathcal{S}}}\left(X_{O \setminus T} \mid \mathrm{do}(X_T = x_T)\right) \\ &= \mathrm{P}_M\left(X_{O \setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S}\right); \end{aligned}$$

(3) for any $T_1 \subseteq V \setminus \mathrm{Anc}_{G^a(M)}(S)$ *and* $x_{T_1} \in \mathcal{X}_{T_1}$, *and any* $T_2 \subseteq (V \setminus \mathrm{Anc}_{G^a(M)}(S))'$ *and* $x_{T_2} \in \mathcal{X}_{T_2}$,

$$\begin{aligned} &\mathrm{P}_{\left(M_{|X_S \in \mathcal{S}}\right)^{\mathrm{twin}}}(X_{(O \cup O') \setminus (T_1 \cup T_2)} \mid \\ &\qquad\qquad \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2})) \\ &= \mathrm{P}_{M^{\mathrm{twin}}}(X_{(O \cup O') \setminus (T_1 \cup T_2)} \mid \\ &\qquad\qquad \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2}), X_S \in \mathcal{S}). \end{aligned}$$

Proof We first prove (1) of Theorem 19. Let $g : \mathcal{X}_W \to \mathcal{X}_V$ be the essentially unique solution function of M. Write $O := V \setminus S$ and $B := \mathrm{Anc}_{G^a(M)}(S)$ and $\star_W = B \cap W$. First note that the function $\hat{g} : \mathcal{X}_{W \setminus B} \times \mathcal{X}_{\star_W} \to \mathcal{X}_{V \setminus S}$ with

$$\hat{g}(x_{W \setminus B}, x_{\star_W}) := g_O(x_{W \setminus B}, x_{\star_W})$$

is the essentially unique solution function of $M_{|\mathcal{S}}$. In fact, for $\mathrm{P}(X_W)$-a.a. $x_W \in \mathcal{X}_W$ and all $x_V \in \mathcal{X}_V$

$$\left\{ \begin{array}{l} x_S = g_S(x_W) \\ x_O = g_O(x_W) \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} x_S = f_S\left(x_V, x_W\right) \\ x_O = f_O\left(x_V, x_W\right) \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} x_S = g^S\left(x_O, x_W\right) \\ x_O = f_O\left(x_V, x_W\right) \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} x_S = g^S\left(x_O, x_W\right) \\ x_O = f_O\left(x_O, g^S(x_O, x_W), x_W\right). \end{array} \right.$$

Let $\hat{\mathrm{P}}$ denote the exogenous probability distribution of $M_{|\mathcal{S}}$, that is, $\hat{\mathrm{P}} := \mathrm{P}(X_{W \setminus B}) \otimes \hat{\mathrm{P}}(X_{\star_W})$, where $\hat{\mathrm{P}}(X_{\star_W}) = \mathrm{P}_M(X_{W \cap B} \mid X_S \in \mathcal{S})$. Recall that we have

$$\mathrm{P}_M(X_V) = g_*\left(\mathrm{P}(X_W)\right)(X_V) \text{ i.e. } \mathrm{P}_M(X_V \in A) = \mathrm{P}(X_W \in g^{-1}(A))$$

for any measurable subset $A \subseteq \mathcal{X}_V$. Then we have for any measurable subset $A \subseteq \mathcal{X}_V$

$$\begin{aligned}
\mathrm{P}_{M_{|\mathcal{S}}}(X_O \in A) &= \hat{\mathrm{P}}\left(X_{\hat{W}} \in \hat{g}^{-1}(A)\right) \\
&= \hat{\mathrm{P}}\left(X_{\hat{W}} \in g_O^{-1}(A)\right) \\
&= \mathrm{P}_M(X_W \in g_O^{-1}(A) \mid X_S \in \mathcal{S}) \\
&= \mathrm{P}_M\left(X_O \in A \mid X_S \in \mathcal{S}\right).
\end{aligned}$$

We then show (2) of Theorem 19. Lemma 17 gives that $\left(M_{|\mathcal{S}}\right)_{\mathrm{do}(X_T = x_T)} = \left(M_{\mathrm{do}(X_T = x_T)}\right)_{|\mathcal{S}}$ for any $T \subseteq V \setminus B$ and $x_T \in \mathcal{X}_T$. We then have for $T \subseteq V \setminus B$ and $x_T \in \mathcal{X}_T$

$$\begin{aligned}
\mathrm{P}_{M_{|\mathcal{S}}}\left(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T)\right) &= \mathrm{P}_{(M_{|\mathcal{S}})_{\mathrm{do}(X_T = x_T)}}\left(X_{O\setminus T}\right) \\
&= \mathrm{P}_{(M_{\mathrm{do}(X_T = x_T)})_{|\mathcal{S}}}\left(X_{O\setminus T}\right) \\
&= \mathrm{P}_{M_{\mathrm{do}(X_T = x_T)}}(X_{O\setminus T} \mid \tilde{g}_S(X_W) \in \mathcal{S}) \\
&= \mathrm{P}_M(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), g_S(X_W) \in \mathcal{S}) \\
&= \mathrm{P}_M(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S}),
\end{aligned}$$

where $\tilde{g}$ is the essentially unique solution function of $M_{\mathrm{do}(X_T = x_T)}$, which satisfies $\tilde{g}_S(x_W) = g_S(x_W)$ for $\mathrm{P}(X_W)$-a.a. $x_W \in \mathcal{X}_W$.

We finally show (3) of Theorem 19. By the definition of conditioning operation and twinning operation, we have $\left((M^{\mathrm{twin}})_{|\mathcal{S}}\right)_{|\mathcal{S}'} = \left((M^{\mathrm{twin}})_{|\mathcal{S}}\right)_{\setminus\mathcal{S}'} = (M_{|\mathcal{S}})^{\mathrm{twin}}$, where $\mathcal{S}' \subseteq \mathcal{X}_{S'}$ is such that $\mathcal{S}' = \mathcal{S}$ and S' is the copy of S. We have from (2) of Theorem 19 that for any $T_1 \subseteq V \setminus \mathrm{Anc}_{G^a(M)}(S)$ and $x_{T_1} \in \mathcal{X}_{T_1}$, and for any $T_2 \subseteq (V \setminus \mathrm{Anc}_{G^a(M)}(S))'$ and $x_{T_2} \in \mathcal{X}_{T_2}$,

$$\begin{aligned}
&\mathrm{P}_{\left(M_{|\mathcal{S}}\right)^{\mathrm{twin}}}\left(X_{(O\cup O')\setminus(T_1\cup T_2)} \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2})\right) \\
&= \mathrm{P}_{\left((M^{\mathrm{twin}})_{|\mathcal{S}}\right)_{\setminus\mathcal{S}'}}\left(X_{(O\cup O')\setminus(T_1\cup T_2)} \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2})\right) \\
&= \mathrm{P}_{(M^{\mathrm{twin}})_{|\mathcal{S}}}\left(X_{(O\cup O')\setminus(T_1\cup T_2)} \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2})\right) \\
&= \mathrm{P}_{M^{\mathrm{twin}}}(X_{(O\cup O')\setminus(T_1\cup T_2)} \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2}), X_S \in \mathcal{S}).
\end{aligned}$$

B.5 Proof of Corollary 20

Corollary 20 (Iterative conditioning) *Assume Assumption 9 with $S = S_1 \cup S_2$ and $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$ with $\mathcal{S}_1 \subseteq \mathcal{X}_{S_1}$ and $\mathcal{S}_2 \subseteq \mathcal{X}_{S_2}$ both measurable. Then $(M_{|\mathcal{S}_1})_{|\mathcal{S}_2}$, $(M_{|\mathcal{S}_2})_{|\mathcal{S}_1}$, and $M_{|\mathcal{S}_1\times\mathcal{S}_2}$ are counterfactually equivalent w.r.t. $V \setminus \mathrm{Anc}_{G^a(M)}(S_1 \cup S_2)$.*[9]

Proof Write $O := V\setminus(S_1\cup S_2)$. From (3) of Theorem 19, it is easy to see that for any $T_1 \subseteq V\setminus\mathrm{Anc}_{G^a(M)}(S_1\cup S_2)$ and $x_{T_1} \in \mathcal{X}_{T_1}$, and any $T_2 \subseteq (V \setminus \mathrm{Anc}_{G^a(M)}(S_1 \cup S_2))'$ and $x_{T_2} \in \mathcal{X}_{T_2}$,

$$\begin{aligned}
&\mathrm{P}_{\left((M_{|\mathcal{S}_1})_{|\mathcal{S}_2}\right)^{\mathrm{twin}}}\left(X_{(O\cup O')\setminus(T_1\cup T_2)} \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2})\right) \\
&= \mathrm{P}_{\left(M_{|\mathcal{S}_1}\right)^{\mathrm{twin}}}\left(X_{(O\cup O')\setminus(T_1\cup T_2)} \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2}), X_{S_2} \in \mathcal{S}_2\right) \\
&= \mathrm{P}_{M^{\mathrm{twin}}}(X_{(O\cup O')\setminus(T_1\cup T_2)} \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2}), X_{S_1} \in \mathcal{S}_1, X_{S_2} \in \mathcal{S}_2) \\
&= \mathrm{P}_{\left(M_{|\mathcal{S}_2}\right)^{\mathrm{twin}}}\left(X_{(O\cup O')\setminus(T_1\cup T_2)} \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2}), X_{S_1} \in \mathcal{S}_1\right) \\
&= \mathrm{P}_{\left((M_{|\mathcal{S}_2})_{|\mathcal{S}_1}\right)^{\mathrm{twin}}}\left(X_{(O\cup O')\setminus(T_1\cup T_2)} \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2})\right).
\end{aligned}$$

Also note that

$$\begin{aligned}
&\mathrm{P}_{\left(M_{|\mathcal{S}_1\times\mathcal{S}_2}\right)^{\mathrm{twin}}}\left(X_{(O\cup O')\setminus(T_1\cup T_2)} \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2})\right) \\
&= \mathrm{P}_{M^{\mathrm{twin}}}(X_{(O\cup O')\setminus(T_1\cup T_2)} \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_2} = x_{T_2}), X_{S_1} \in \mathcal{S}_1, X_{S_2} \in \mathcal{S}_2).
\end{aligned}$$

[9] See Definition 33 or Bongers et al. (2021, Definition 4.5) for the definition of counterfactual equivalence.

B.6 Proof of Proposition 22

Proposition 22 (Conditioning & marginalization) *Assume Assumption 9 and let $L \subseteq V \setminus S$. Then we have $(M_{\setminus L})_{|\mathcal{S}} = (M_{|\mathcal{S}})_{\setminus L}$.*

Proof When conditioning, we merge the exogenous ancestors of S, which commutes with marginalization. It is easy to see that updating the exogenous probability distribution also commutes with marginalization. Also note that $(M_{\setminus L})_{\setminus S} = (M_{\setminus S})_{\setminus L}$, since $S \cap L = \emptyset$ (Bongers et al., 2021, Proposition 5.4). Combining these three implies that $(M_{\setminus L})_{|\mathcal{S}} = (M_{|\mathcal{S}})_{\setminus L}$.

Supplement C. Details of Some Examples

C.1 Example of Definition 13

Here we show an example of the purely graphical conditioning operation, i.e., Definition 13. Assume we are given a graph G as shown in Figure 10. Then conditioning on the node V_5 gives the graph $G_{|V_5}$ shown in Figure 10.

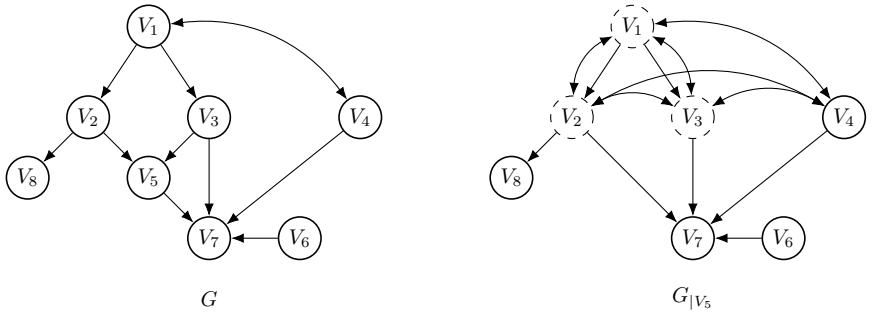

Figure 10: DMG G and its Conditioned DMG $G_{|V_5}$.

C.2 Example 27

We explain why one has two different answers to the same question based on G and $\tilde{G}$, respectively. For an SCM with graph G, one has:

$$\mathrm{TCE}(Y; c' \to c) := \mathbb{E}[F \mid \mathrm{do}(C = c)] - \mathbb{E}[F \mid \mathrm{do}(C = c')] = \mathbb{E}[F \mid C = c] - \mathbb{E}[F \mid C = c'].$$

On the other hand, for an SCM with graph $\tilde{G}$, one has:

$$\begin{aligned}
\mathrm{TCE}(Y; c' \to c) &:= \mathbb{E}[F \mid \mathrm{do}(C = c)] - \mathbb{E}[F \mid \mathrm{do}(C = c')] \\
&= \sum_a \left(\mathbb{E}[F \mid C = c, A = a] - \mathbb{E}[F \mid C = c', A = a]\right) \mathrm{P}(A = a) \\
&\neq \sum_a \left(\mathbb{E}[F \mid C = c, A = a]\mathrm{P}(A = a \mid C = c) - \mathbb{E}[F \mid C = c', A = a]\mathrm{P}(A = a \mid C = c')\right) \quad \text{(in general)} \\
&= \mathbb{E}[F \mid C = c] - \mathbb{E}[F \mid C = c'].
\end{aligned}$$