# Foundations of Structural Causal Models with Latent Selection

Leihao Chen[*], Onno Zoeter[†], and Joris M. Mooij[‡]

November 18, 2025

**Abstract**

Three distinct phenomena complicate statistical causal analysis: latent common causes, causal cycles, and latent selection. Foundational works on Structural Causal Models (SCMs), e.g., Bongers et al. (2021, Ann. Stat., 49(5): 2885-2915), treat cycles and latent variables, while an analogous account of latent selection is missing. The goal of this article is to develop a theoretical foundation for modeling latent selection with SCMs. To achieve that, we introduce a conditioning operation for SCMs: it maps an SCM with explicit selection mechanisms to one without them while preserving the causal semantics of the selected subpopulation. Graphically, in Directed Mixed Graphs we extend bidirected edges—beyond latent common causes—to also encode latent selection. We prove that the conditioning operation preserves simplicity, acyclicity, and linearity of SCMs, and interacts well with marginalization, conditioning, and interventions. These properties make those three operations valuable tools for causal modeling, reasoning, and learning after abstracting away latent details (latent common causes and selection). Examples show how this abstraction streamlines analysis and clarifies when standard tools (e.g., adjustment, causal calculus, instrumental variables) remain valid under selection bias. We hope that these results deepen the SCM-based understanding of selection bias and become part of the standard causal modeling toolbox to build more reliable causal analysis.

**Keywords:** Causal Model Abstraction, Causal Modeling, Conditioning Operation, Graphical Models, Selection Bias, Structural Causal Models

[*]Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, the Netherlands; l.chen2@uva.nl

[†]Booking.com, The Netherlands; onno.zoeter@booking.com

[‡]Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, the Netherlands; j.m.mooij@uva.nl

# Contents

# 1 Introduction

Bongers et al. (2021) provide a general measure-theoretic foundational theory for causal modeling with Structural Causal Models (SCMs) with cycles and latent variables, but an analogous treatment for latent selection is still absent. Addressing (latent) selection bias remains a significant challenge (Wald, 1943; Heckman, 1979; Zhao et al., 2021; Fryer Jr, 2019; Cooper, 1995). For example, in some cases unconsciously selecting samples can induce "spurious dependency" among collected samples, and therefore the famous Berkson's paradox arises (Berkson, 1946; Munafo et al., 2016). There are many types of selection bias without universally accepted definitions and various methods to address them (Lu et al., 2022; Smith, 2020). In this work, we focus on "truncating selection bias," which occurs when an underlying (unobserved) filtering process, denoted "$X_S \in \mathcal{S}$", selects individual samples where the variable $X_S$ takes values within a set $\mathcal{S}$. In probability theory, this can be modeled as conditioning on the event $\{X_S \in \mathcal{S}\}$.

To understand its structural behavior, one approach is to model selection bias via a causal model that explicitly describes the selection mechanism (Pearl, 2009; Bareinboim and Pearl, 2012; Daniel et al., 2012; Abouei et al., 2024a; Hernán et al., 2004). This necessitates detailed knowledge of the selection mechanism. However, in many situations, the selection mechanism is unobserved (Cooper, 1995), which makes such knowledge unavailable and introduces a layer of complexity with infinitely many possibilities. The goal of the current work is to study how to model latent selection by effectively abstracting away its details in an SCM (Pearl, 2009; Bongers et al., 2021).

## 1.1 Motivation

Marginalization of causal models is a powerful tool for abstracting away latent details, which makes causal modeling more manageable and trustworthy (Bongers et al., 2021; Pearl, 2009; Evans, 2016). By marginalizing out latent variables, we use one simplified model to represent infinitely many complex models, abstracting away unnecessary latent details while preserving essential causal information such as observational/interventional/counterfactual distributions, $d$-separations or $\sigma$-separations (Pearl, 2009; Forré and Mooij, 2017), and ancestral relationships among the observed variables. The SCM marginalization and causal graph marginalization interact well and part of the nice properties can be compactly expressed via Figure 1.[1]

For instance, the model $G$ in Figure 2 effectively abstracts models $G^i$ for $i = 1, \ldots, 5, \ldots$, yielding the same identification result (under discreteness and positivity assumptions) regardless of the latent structure (front-door criterion (Pearl, 2009)):[2]

$$\mathrm{P}(C = c \mid \mathrm{do}(S = s)) = \sum_t \mathrm{P}(C = c \mid T = t)\mathrm{P}(T = t \mid S = s).$$

[1] Replacing $M$ with the intervened model $M_{\mathrm{do}(X_T = x_T)}$ where we intervene the variable $X_T$ to take on the value $x_T$ (cf. Definition 2.2), we get the corresponding results for interventional distributions and intervened graphs.

[2] More generally, the ID-algorithm was first proved to be sound and complete for models with bidirected edges and then the results can be translated to the case with arbitrary latent structures via marginalization (Tian and Pearl, 2002; Huang and Valtorta, 2008; Richardson et al., 2023).

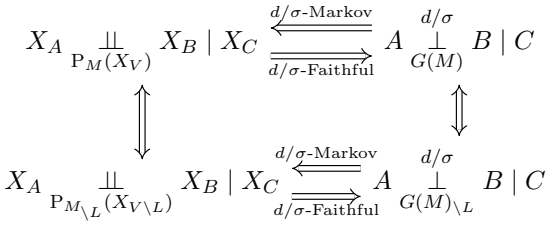



Figure 1: The logical relations between stochastic conditional independence in simple SCM $M$ and marginalized model $M_{\setminus L}$, and graphical separation ($d$- or $\sigma$-separation) in causal graph $G(M)$ and marginalized graph $G(M)_{\setminus L}$. SCM $M$ has endogenous variables $X_V$ with $X_L$ latent, and $A, B, C \subseteq V \setminus L$. The terms "$d/\sigma$-Markov" and "$d/\sigma$-Faithful" represent $d$- or $\sigma$-Markov property and $d$- or $\sigma$-faithfulness, respectively, regarding $\mathrm{P}_M(X_V)$ and $G(M)$ (top), and $\mathrm{P}_{M_{\setminus L}}(X_{V\setminus L})$ and $G(M)_{\setminus L}$ (bottom).

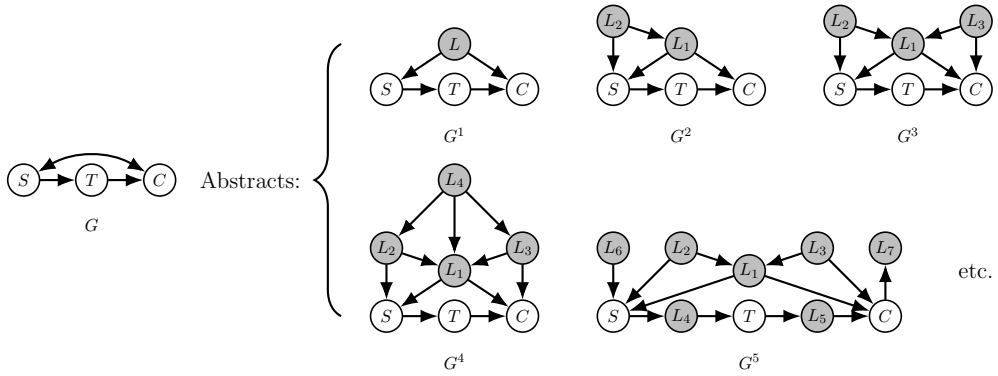



Figure 2: $G$ effectively abstracts $G^i$ for $i = 1, \ldots, 5, \ldots$

**Motivating questions** Selection bias is ubiquitous, often latent, and can lead to biased results; therefore, not taking it into account may lead to an untrustworthy model. Unfortunately, marginalization is not able to deal with latent selection bias (cf. Example 1.1). Considering this, the following questions arise naturally:

Q1 Given an SCM $M$ with a selection mechanism $X_S \in \mathcal{S}$, can we always find an SCM without a selection mechanism to faithfully represent $(M, X_S \in \mathcal{S})$?[3] (cf. Appendix D.1)

Q2 If not, which part of the causal semantics of $(M, X_S \in \mathcal{S})$ can be represented by an SCM in general? Can we construct a transformation that transforms $(M, X_S \in \mathcal{S})$ into an SCM $M_{|X_S \in \mathcal{S}}$ so that $M_{|X_S \in \mathcal{S}}$ encodes this part of the causal semantics? What properties does it have? (cf. Theorem 3.14, Definition 3.5, Section 3.2.2, and Proposition D.1)

[3]Note that this question is trickier than it seems. Finding an SCM without a selection mechanism to represent another SCM with a selection mechanism is, to some extent, analogous to the problem of finding a DAG to represent the marginalized model of another DAG, which is impossible in general (Richardson and Spirtes, 2002). See also Blom et al. (2020) for some data-generating processes with clear causal interpretation that cannot be modeled by SCMs.

Q3 Can we similarly construct a transformation on causal graphs such that it is compatible with the transformation at the level of SCMs? What properties does it have and what is the relation between the "conditioned SCM" $M_{|X_S \in \mathcal{S}}$ and the "conditioned causal graph" $G(M)_{|S}$ (Figure 3)? (cf. Definition 3.23, Section 3.3)

We answer these questions in the current manuscript.

$$\begin{array}{ccc} X_A \underset{\mathrm{P}_M(X_V)}{\perp\!\!\!\perp} X_B \mid X_C, X_S \in \mathcal{S} & \overset{?}{\underset{?}{\rightleftharpoons}} & A \overset{d/\sigma}{\underset{G(M)}{\perp}} B \mid C \cup S \\ \Big\Updownarrow ? & & \Big\Updownarrow ? \\ X_A \underset{\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_{V \setminus S})}{\perp\!\!\!\perp} X_B \mid X_C & \overset{?}{\underset{?}{\rightleftharpoons}} & A \overset{d/\sigma}{\underset{G(M)_{|S}}{\perp}} B \mid C \end{array}$$

Figure 3: What are the relations between stochastic conditional independence in $M_{|X_S \in \mathcal{S}}$ and graphical separation in $G(M)_{|S}$? The answer is shown in Figure 11.

**A motivating example** To illustrate, we first discuss a toy example, demonstrating that marginalization is *not* appropriate for abstracting away selection bias and how to obtain correct results without assuming any specific details about the latent selection mechanism.

**Example 1.1** (Car mechanic)**.** A car starts successfully if its battery is charged and its start engine is operating. Introduce latent binary endogenous variables $B_0$ ("battery"), $E_0$ ("start engine") and $S_0$ ("car starts") measured at time $t_0$ and observed variables $B_1, E_1$ and $S_1$ with a similar meaning for the same car but measured at time $t_1$ with $t_1 > t_0$. We model this[4] by the following SCM $M$ and denote by $M^*$ its marginalized model on observed endogenous variables $B_1$, $E_1$, and $S_1$.

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

where $U_B$ and $U_E$ are latent exogenous independent Bernoulli-distributed random variables with parameters $1-\delta$ and $1-\epsilon$. Their graphs are shown in Figure 4.

The question is whether there exists an SCM with variables $B_1, E_1, S_1$ encoding the causal semantics of $M$ for the subpopulation of cars for which $S_0 = 0$. Consider the SCM $\widetilde{M}$, whose graph is depicted in Figure 4, given by

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

| $\widetilde{\mathrm{P}}(U_B, U_E)$ | $U_E = 0$ | $U_E = 1$ |
|---|---|---|
| $U_B = 0$ | $\frac{\delta\epsilon}{\delta+(1-\delta)\epsilon}$ | $\frac{\delta(1-\epsilon)}{\delta+(1-\delta)\epsilon}$ |
| $U_B = 1$ | $\frac{(1-\delta)\epsilon}{\delta+(1-\delta)\epsilon}$ | $0$ |

[4] For illustration, we assume such a simplified model. One can add more (endogenous or exogenous random) variables to the model.

As one can check,

$$\begin{aligned}
\mathrm{P}_{\widetilde{M}}(B_1, E_1, S_1) &= \mathrm{P}_M(B_1, E_1, S_1 \mid S_0 = 0) \neq \mathrm{P}_{M^*}(B_1, E_1, S_1), \\
\mathrm{P}_{\widetilde{M}}(S_1 = 1 \mid \mathrm{do}(B_1 = 1)) &= \mathrm{P}_M(S_1 = 1 \mid \mathrm{do}(B_1 = 1), S_0 = 0) \neq \mathrm{P}_{M^*}(S_1 = 1 \mid \mathrm{do}(B_1 = 1)), \\
\mathrm{P}_{\widetilde{M}}(S_1 = 1 \mid \mathrm{do}(E_1 = 1)) &= \mathrm{P}_M(S_1 = 1 \mid \mathrm{do}(E_1 = 1), S_0 = 0) \neq \mathrm{P}_{M^*}(S_1 = 1 \mid \mathrm{do}(E_1 = 1)).
\end{aligned}$$

The car mechanic is only interested in cars that failed to start at an early time $t_0$ and are sent to the studio at a later time $t_1$. 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 the marginalized model $M^*$ does not possess the correct causal semantics of the subpopulation. Furthermore, the graph $G(\widetilde{M})$ correctly expresses that $B_1$ and $E_1$ might be dependent in the subpopulation (given $S_0 = 0$) via the $d$-separation criterion for acyclic directed mixed graphs (Richardson, 2003), while the graph $G(M^*)$ wrongly claims that $B_1$ and $E_1$ are independent. Therefore, $\widetilde{M}$ effectively abstracts away irrelevant latent modeling details: (i) the latent variables $B_0$, $E_0$ and $S_0$, (ii) their causal mechanisms, and (iii) the explicit selection step on $S_0 = 0$. However, the marginalized model $M^*$ does not, which shows that marginalization alone cannot abstract latent selection mechanisms.

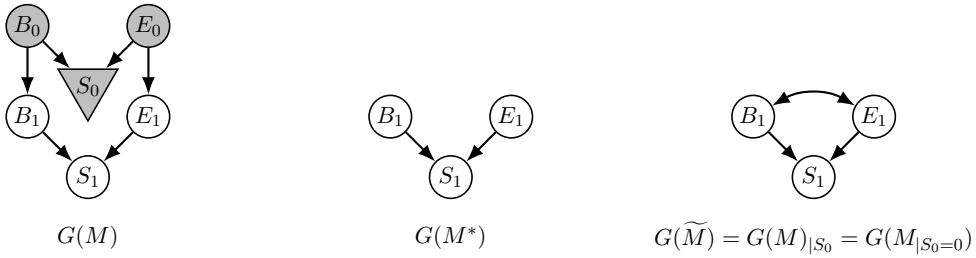



Figure 4: The causal graphs of the SCMs $M$, $M^*$ and $\widetilde{M}$ in Example 1.1. The gray nodes are latent and the triangle means conditioning on $S_0$ to take some specific values (cf. Notation 2.3, Definition 3.2). Marginalizing out all the latent variables yields $M^*$, while conditioning out $S_0$ (cf. Definition 3.5) and marginalizing out the remaining latent variables yields $\widetilde{M}$.

Note that in Example 1.1, we can obtain the model $\widetilde{M}$ directly from $M$ (cf. Definition 3.5 and Example 3.8), by (i) merging $U_B$ and $U_E$, (ii) replacing $\mathrm{P}_M(U_B, U_E)$ by $\mathrm{P}_M(U_B, U_E \mid S_0 = 0)$, and (iii) 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). In fact, this procedure can be generalized and gives the desired transformation of Question Q2, as we show in Section 3.

## 1.2 Our contribution

Pearl (2009, p.163) claims that (when doing causal modeling):"...bidirected arcs should be assumed to exist, by default, between any two nodes in the diagram. They should be deleted only by well-motivated justifications, such as the unlikely existence of a common cause for the two variables *and the unlikely existence of selection bias.*" Although marginalization makes

it clear how bidirected edges can represent latent common causes, a rigorous approach to representing (latent) selection bias with bidirected edges has not been formalized yet.

Our main contribution is that we provide a rigorous approach to representing *(latent) selection bias* with *bidirected edges*. To be more precise, given a Structural Causal Model $M$ with a selection mechanism $X_S \in \mathcal{S}$ (cf. Definition 3.2) where variable $X_S$ takes values in a measurable subset $\mathcal{S}$, we define a transformation that maps $(M, X_S \in \mathcal{S})$ to a "conditioned" SCM $M_{|X_S \in \mathcal{S}}$ without any accompanying selection mechanism, so that $M_{|X_S \in \mathcal{S}}$ is an *effective abstraction* of $M$ w.r.t. the selection $X_S \in \mathcal{S}$ in the sense that:

(i) the conditioned SCM $M_{|X_S \in \mathcal{S}}$ correctly encodes as much *causal semantics* (observational, interventional and counterfactual) of $M$ of the *subpopulation* $X_S \in \mathcal{S}$ as possible (cf. Theorems 3.14 and 3.17 and Section D.1);

(ii) the conditioning operation preserves important model classes, e.g., linear, acyclic and simple SCMs (cf. Proposition 3.10);

(iii) this conditioning operation interacts well with other operations on SCMs, e.g., intervention, marginalization, and the conditioning operation itself (cf. Lemma 3.13, Propositions 3.18 and 3.19);

(iv) one can read off *qualitative causal information* about $M$ under the selected subpopulation from the *causal graph* of $M_{|X_S \in \mathcal{S}}$ and the *conditioned graph* $G(M)_{|S}$ (cf. Definition 3.23, Theorems A.8 and A.11, Corollary 3.31).

In Section 3, we will introduce the rigorous mathematical definition of the conditioning operation (Definition 3.5) and demonstrate that it possesses all the aforementioned properties.

The significance of this conditioning operation lies in the fact that we can take $M_{|X_S \in \mathcal{S}}$ as a simplified "proxy" for $M$ w.r.t. the selection $X_S \in \mathcal{S}$, which effectively abstracts away details about latent selection (i.e., satisfying the properties listed previously). This makes it a versatile tool for causal inference tasks with latent selection bias. Specifically:

(1) Causal Reasoning: One can directly apply all the causal inference tools for SCMs on $M_{|X_S \in \mathcal{S}}$, e.g., identification results (adjustment criterion and Pearl's do-calculus), ID-algorithm and instrumental inequality, which simplifies causal reasoning tasks under latent selection bias (cf. Examples 4.3, 4.6 and 4.8).

(2) Causal Modeling: Utilizing the marginalization and conditioning operation, we can represent infinitely many SCMs with a single marginalized conditioned SCM. This significantly streamlines causal modeling, eliminating the need to enumerate all possibilities with different latent structures and selection mechanisms. Moreover, it improves the robustness and trustworthiness of the model by reducing the sensitivity to various causal assumptions (cf. Examples 4.1 and 4.12).

(3) Causal Discovery: Many algorithms exclude selection bias by assumption—an often unrealistic idealization (Cooper, 1995). Nevertheless, methods originally designed for latent common causes without selection bias can, under suitable conditions, be applied directly to selected data, with their outputs interpreted as learning the conditioned

model $M_{|X_S \in \mathcal{S}}$ (cf. Example 4.7). This requires no redesign of the algorithm: the unmodified procedure still admits certain causal interpretation of its output in the presence of selection.[5]

It is worth mentioning that many of our results rely on two facts: bidirected edges in DMGs can be used to represent latent selection bias, and DMGs admit an interpretation as causal graphs of SCMs. There is one subtlety, though: not all the endogenous variables of the conditioned SCM retain their causal interpretation. Interventions targeting such nodes yield predictions that are typically incompatible with those of the corresponding interventions on the original SCM in the presence of the selection mechanism (see also Section 3.4). However, these "non-intervenable" endogenous variables are easily identified as the ancestors of the selection variables.

## 1.3 Connections to related work

In a series of papers (Bareinboim and Pearl, 2012; Bareinboim and Tian, 2015), the authors explored the 's-recoverability' problem, aiming to recover causal quantities for the whole population from selected data. This investigation operated under qualitative causal assumptions on the selection nodes, explicitly expressed in terms of causal graphs. However, such knowledge about selection nodes is not always available (Richardson and Robins, 2013a, Footnote 11). In the current work, we focus on the problem of how to *model selection bias with an SCM without explicitly modeling the selection mechanism* and draw (causal) conclusions for the selected subpopulation.

There are graphical models with well-behaved marginalization and conditioning operations such as maximal ancestral graphs (MAGs) (Richardson and Spirtes, 2002), $d$-connection graphs (Hyttinen et al., 2014) and $\sigma$-connection graphs (Forré and Mooij, 2018). Among them, MAGs were originally developed as a model class representing the conditional independence models of the marginalized conditioned conditional independence models of DAGs. By summarizing the common causal features of causal DAGs represented by a MAG, one can give a causal interpretation to MAGs and call them causal MAGs. One single causal MAG can represent infinitely many SCMs with different graphs but the same conditional independences among observed variables. Interpreting a graph as a causal graph of an SCM and as a causal MAG respectively will not give the same causal conclusions in general.[6] Due to the nature of model abstraction, MAGs are well suited for causal discovery, and one can further draw *some* causal conclusions from MAGs (Spirtes et al., 1995; Richardson, 2003; Zhang, 2008; Mooij and Claassen, 2020). However, MAGs are not always suitable for *causal modeling* under selection bias in some cases, since: (i) it is not clear how to read off causal relationships (direct causal relations, confounding) from MAGs; (ii) there are no causal identification results for MAGs under selection bias and causal cycles yet; (iii) currently the standard theory of MAGs cannot deal with counterfactual reasoning. On the other hand, our conditioning operation transforms an SCM with selection mechanisms to an ordinary SCM, which carries an intuitive

[5] Due to the abstraction nature of conditioned SCM, we lose some information in this process and see Sections 3.4 and D.1 and Theorems 3.14 and 3.17 for the subtlety of causal interpretation of conditioned SCM.

[6] For example, consider a graph consisting of $A \to B \to C$ and $A \to C$. If it is a causal graph of an SCM, then we can conclude that variable $A$ has a direct causal effect on $C$ according to this model and we can identify $\mathrm{P}(C = c \mid \mathrm{do}(A = a)) = \mathrm{P}(C = c \mid A = a)$ under the positivity and discreteness assumption. However, if it is a MAG, then we cannot obtain the above two conclusions.

causal interpretation. All the theory for SCMs (causal identification, cycles, counterfactual reasoning) can be directly extended to the case with selection bias via the conditioning operation. Therefore, our results can address causal inference tasks such as fairness analysis (Kusner et al., 2017; Zhang and Bareinboim, 2018), causal modeling of dynamical systems (Bongers et al., 2022; Peters et al., 2022) and biological systems with feedback loops (Versteeg et al., 2022) under selection bias (cf. Definition 2.1, Remarks 4.5 and 4.9, Examples 4.6, 4.10 and 4.12). Another subtle difference between SCM conditioning and MAG conditioning is that they consider different forms of conditioning (cf. Example B.6).

Although causal graphs provide a means to differentiate selection bias from confounding due to common causes (Hernán et al., 2004; Cooper, 1995), the potential outcome community tends to amalgamate the two (Richardson and Robins, 2013a; Hernán and Robins, 2020). In many cases, one can be sure about the existence of "non-causal dependency", but cannot be sure whether it is induced by a latent common cause or latent selection bias or the combination of the two (see e.g., Richardson and Robins (2013a, Footnote 11) and Pearl (2009, p.163)). Our conditioning operation formalizes this ambiguity within SCMs. Graphically, we employ bidirected edges to symbolize the dependence of two variables arising from either unmeasured common causes, latent selection bias, or any intricate combination of the two. Therefore, in causal modeling, our work allows the modeler to be able to represent such non-causal dependency abstractly via bidirected edges.

Some work considers the abstraction of causal models from the perspective of grouping low-level variables to high-level variables and merging values of variables (Rubenstein et al., 2017; Beckers and Halpern, 2019). Geiger et al. (2023) study the so-called "constructive abstraction" of causal models. They show that it can be characterized as a composition of clustering sets of variables, merging values of variables, and marginalization. Our conditioning operation does not fall under the umbrella of "constructive abstraction" of Geiger et al. (2023).

## 1.4 Outline

In Section 2, we review basic notions of SCMs and fix the notation used throughout the article; additional preliminaries are deferred to Section A to save space. In Section 3.1, we give a formal definition of SCMs with selection mechanisms. In Sections 3.2 and 3.3, we introduce the conditioning operations for SCMs and DMGs, respectively, and study their mathematical properties and mutual relations; Theorems 3.14 and 3.25 and Proposition 3.29 contain the main results. In Section 3.4, we discuss important caveats concerning the interpretation of conditioned SCMs. Further remarks and examples related to Section 3 are collected in Section B, while all proofs of the results in Section 3 are provided in Section C.

We illustrate the applicability of the conditioning operation through a series of examples in Section 4, including generalized versions of Reichenbach's principle, the back-door theorem, the ID-algorithm, instrumental variables, causal model learning, mediation analysis, and a real-world causal modeling exercise on COVID-19. In Section D.1, we show that SCMs without selection mechanisms are, in general, not flexible enough to represent SCMs with selection mechanisms, which answers Questions Q1–Q3 in combination with the discussion in Section 3. Finally, in Sections D.2 and D.3, we explore alternative conditioning operations, including variants based on different decompositions of exogenous variables, a conditioning operation for causal Bayesian networks, and a conditioning operation for SCMs with exogenous non-stochastic input variables (cf. Definition D.9).

# 2 Preliminaries and notation

This section provides the necessary background on SCMs and introduces the notions of common cause and confounding. To save space, additional preliminaries on SCMs are deferred to Section A. We follow the formal setup of Bongers et al. (2021), which allows us to formulate the theory for "simple" SCMs, a class that includes acyclic SCMs as well as well-behaved cyclic SCMs. However, we also depart from Bongers et al. (2021) in several respects—for instance, we allow for non-intervenable variables and introduce new node types (dashed nodes and triangle nodes). In this section we define (i) SCMs, (ii) (hard) interventions, (iii) SCM solution functions, (iv) simple SCMs, (v) marginalization, and (vi) basic causal relationships such as common cause and confounding (cf. Definitions 2.1, 2.2, 2.4, 2.6, 2.7 and 2.12). We also fix notation for causal graphs and (conditional) interventional distributions (cf. Notations 2.3 and 2.11).

## 2.1 Structural Causal Model (SCM)

**Definition 2.1** (Structural Causal Model)**.** *A **Structural Causal Model (SCM)** is a tuple $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ such that*

*(i) $V, W$ are disjoint finite sets of labels for the **endogenous variables** and the **latent exogenous random variables**, respectively;*

*(ii) the **state space** $\mathcal{X} = \prod_{i \in V \dot\cup W} \mathcal{X}_i$ is a product of standard measurable spaces $\mathcal{X}_i$;*

*(iii) the **exogenous distribution** $\mathrm{P}$ 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 $\mathcal{X}_w$;*

*(iv) the **causal mechanism** is specified by the measurable mapping $f : \mathcal{X} \to \mathcal{X}_V$.*

**Definition 2.2** (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***

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

This replaces the targeted endogenous variables with 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. If an endogenous variable is modeled as "intervenable", then we say that we model it as causal or that it has a causal interpretation.[7]

[7] Although some variables are modeled as "non-intervenable", one can mathematically define an intervention on them. However, one should be careful with the causal interpretation (Pearl, 2019, 2015). Similar problems arise in the work of causal model abstraction such as Rubenstein et al. (2017) and Beckers and Halpern (2019). An intervention on the "high-level" variables in the abstracted models may not correspond to a well-defined intervention on the "low-level" variables in the detailed models. One can keep track of the "allowed intervention targets" $\mathcal{I} \subseteq V$ and augment $M$ to $(M, \mathcal{I})$. Similarly, one can also encode the information about which variables are latent or not in the definition of an SCM. This would introduce four types of endogenous nodes, which makes the notation quite heavy. Note that mathematically they can often be treated equally and the difference comes only at the phase of modeling. Therefore, we do not distinguish these nodes in the definition of SCMs and only mark them informally with different types of nodes in the causal graphs (cf. Notation 2.3). Another method is to introduce the so-called regime indicators (Dawid, 2002, 2021) to indicate on the graphs which variables are causal and which are purely probabilistic. This usually makes causal graphs much more inflated and requires us to introduce a conditioning operation for SCMs with exogenous non-stochastic input variables (cf. Definition D.9). To ease notation and the reader's mental burden, we do not adopt this approach, either.

To avoid confusion, most of the time we will only consider interventions $\mathrm{do}(X_T = x_T)$ with intervention target $T$ a subset of the intervenable nodes in $V$. We consider all exogenous random variables as non-intervenable. Other types of interventions can be defined, such as soft or stochastic ones (Correa and Bareinboim, 2020).

Given an SCM $M$, one can define its causal graph $G(M)$ and its augmented causal graph $G^a(M)$ to give intuitive and compact graphical representations of the causal model (see Definition A.3). One can read off useful causal information directly from the causal graph without knowing the details of the underlying SCM.

**Notation 2.3** (Causal graphs)**.** *In all the causal graphs, we use gray nodes to represent latent variables. Dashed nodes represent non-intervenable variables, and solid nodes represent intervenable variables. Exogenous variables are assumed latent and non-intervenable, so they are marked gray and dashed. Triangle nodes mean that there are selection mechanisms conditioning on the corresponding variables to take some specific values.*[8] *We sometimes abuse the notation by identifying the label and random variables in causal graphs.*

**Definition 2.4** (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 $x_{V\setminus S} \in \mathcal{X}_{V\setminus S}$,*[9] *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 2.5** (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 $\widetilde{g}^S$ 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 $g^S(x_{V\setminus S}, x_W) = \widetilde{g}^S(x_{V\setminus S}, x_W)$. If $M$ has an essentially unique solution function w.r.t. $V$, we call it **uniquely solvable**.*

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

**Definition 2.6** (Simple SCMs)**.** *An SCM $M$ is called a **simple SCM** if it is uniquely solvable w.r.t. each subset $S \subseteq V$.*

Note that *all acyclic SCMs are simple* (Bongers et al., 2021, Proposition 3.4). One benefit of introducing the class of simple SCMs is that it preserves the most convenient properties of acyclic SCMs but allows for weak cycles. We focus on simple SCMs in this work so that we can avoid mathematical technicalities and focus on conceptual issues (cf. Assumption 3.1 and Remark B.2).

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 while preserving the causal semantics of the remaining variables (Bongers et al., 2021).

[8] A triangle looks like a funnel, which means that we filter out some samples based on the values of variable $X_S$.

[9] The ordering of the two quantifiers does matter and cannot be changed in the definition. See e.g., Bongers et al. (2021, Lemma F.11) for more details.

**Definition 2.7** (Marginalization). *Let $M$ be a simple SCM, $L \subseteq V$, and $g^L$ be a solution function of $M$ w.r.t. $L$. Then we call $M_{\setminus L} = (V \setminus L, W, \mathcal{X}_{V\setminus L} \times \mathcal{X}_W, \mathrm{P}, \widetilde{f})$ with*

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

*a **marginalization** of $M$ over $L$.*

For SCMs, one can introduce a hierarchy of equivalence relations. Observational, interventional, and counterfactual equivalence mean that two SCMs have the same observational, interventional, and counterfactual semantics, respectively (see Bongers et al. (2021, Definitions 4.1, 4.3 and 4.5) or Definition A.5). Counterfactual equivalence is strictly stronger than interventional equivalence, and interventional equivalence is strictly stronger than observational equivalence. *Equivalence of SCMs* is also an equivalence notion stronger than the three equivalence notions mentioned above.

**Definition 2.8** (Equivalence). *An SCM $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ is **equivalent** to an SCM $\widetilde{M} = (V, W, \mathcal{X}, \mathrm{P}, \widetilde{f})$ if for all $v \in V$, for $\mathrm{P}$-a.a. $x_W \in \mathcal{X}_W$ and for all $x_V \in \mathcal{X}_V$,*

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

*If $M$ and $\widetilde{M}$ are equivalent, we write $M \equiv \widetilde{M}$.*[10]

A simple SCM induces a collection of distributions that includes its observational distribution and interventional distributions. Besides, one can describe counterfactual semantics of an SCM by performing interventions in its twin SCM (see Definition A.1). Potential outcomes are also used to express counterfactual semantics (Hernán and Robins, 2020; Rubin, 1974). We can define potential outcomes via simple SCMs (Bongers et al., 2021).[11]

**Definition 2.9** (Potential outcome). *Let $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ be a simple SCM, $T \subseteq V$ be a subset, and $x_T \in \mathcal{X}_T$ be a value. The potential outcome under the perfect intervention $\mathrm{do}(X_T = x_T)$ is defined as $X_V(x_T) := (g^{V\setminus T}(x_T, X_W), x_T)$, where $g^{V\setminus T} : \mathcal{X}_T \times \mathcal{X}_W \to \mathcal{X}_{V\setminus T}$ is the (essentially unique) solution function of $M$ w.r.t. $V \setminus T$ and $X_W$ is a (fixed) random variable such that $X_W \sim \mathrm{P}$.*

**Definition 2.10** (Potential-outcome equivalence). *We say two SCMs $M^1 = (V, W^1, \mathcal{X}^1, \mathrm{P}^1, f^1)$ and $M^2 = (V, W^2, \mathcal{X}^2, \mathrm{P}^2, f^2)$ are **potential-outcome equivalent** if $\mathcal{X}^1_V = \mathcal{X}^2_V$ and*

$$\mathrm{P}_{M^1}(\{X_V(x_{T_i})\}_{1\le i\le n}) = \mathrm{P}_{M^2}(\{X_V(x_{T_i})\}_{1\le i\le n})$$

*for all $T_i \subseteq V$, all $x_{T_i} \in \mathcal{X}_{T_i}$ and $i = 1, \ldots, n$.*

Now we present the notations of (conditional) interventional distributions that we use in the current manuscript.

[10] For defining equivalence of SCMs, one does not need to assume that $M$ and $\widetilde{M}$ have the same sets of exogenous nodes but only needs the two sets to be isomorphic. For simplicity, we do not specify this in detail.

[11] In most of the potential outcome literature, potential outcomes are taken as primitives and are not induced by an underlying SCM.

**Notation 2.11** ((Conditional) Interventional distributions)**.** *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* $M$*. Let* $S \subseteq V$*,* $O \coloneqq V \setminus S$*, and* $T \subseteq V$ *with* $T \cap S = \emptyset$*. For a measurable subseteq* $\mathcal{S} \subseteq \mathcal{X}_S$*, we use*

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

*to represent the probability distribution of* $X_O$ *when first intervening on* $X_T = x_T$ *and second conditioning on* $X_S \in \mathcal{S}$*, assuming* $\mathrm{P}_{M_{\mathrm{do}(X_T=x_T)}}(X_S \in \mathcal{S}) > 0$*.*[12] *Using the notation of potential outcomes, we have*

$$\mathrm{P}_M(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S}) = \mathrm{P}_M(X_{O\setminus T}(x_T) \mid X_S(x_T) \in \mathcal{S}),$$

*which is not equal to* $\mathrm{P}_M(X_{O\setminus T}(x_T) \mid X_S \in \mathcal{S})$ *in general if* $T \cap \mathrm{Anc}_{G(M)}(S) \neq \emptyset$*. If* $T = \emptyset$*, then* $\mathrm{P}_M(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S}) = \mathrm{P}_M(X_O \mid X_S \in \mathcal{S})$ *and* $X_{O\setminus T}(x_T) = X_O$*.*

## 2.2 Common cause and confounding

"Confounder", "common cause" and "confounding" have diverse and vague meanings in different literature (VanderWeele and Shpitser, 2013). For conceptual clarity, we give formal definitions for these notions in the setting of acyclic SCMs.[13]

**Definition 2.12** (Common cause and confounding)**.** *Let* $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ *be an acyclic SCM and* $A, B, C \in V$ *be distinct intervenable nodes.*

*(1) We say that* $X_C$ *is a* ***common cause*** *of* $X_A$ *and* $X_B$ *according to* $M$ *if there exist* $x_A \in \mathcal{X}_A$*,* $x_B \in \mathcal{X}_B$*,* $x_C \in \mathcal{X}_C$*, and* $x'_C \in \mathcal{X}_C$ *such that*

$$\begin{aligned}&\mathrm{P}_M(X_A \mid \mathrm{do}(X_C = x_C), \mathrm{do}(X_B = x_B)) \neq \mathrm{P}_M(X_A \mid \mathrm{do}(X_B = x_B)) \text{ and} \\ &\mathrm{P}_M(X_B \mid \mathrm{do}(X_C = x'_C), \mathrm{do}(X_A = x_A)) \neq \mathrm{P}_M(X_B \mid \mathrm{do}(X_A = x_A)).\end{aligned}$$

*(2) Assume that for all* $x_B \in \mathcal{X}_B$*, we have* $\mathrm{P}_M(X_A) = \mathrm{P}_M(X_A \mid \mathrm{do}(X_B = x_B))$*. We say that there is* ***confounding bias*** *between* $X_A$ *and* $X_B$*, if there exists a measurable subset* $\mathcal{A} \subseteq \mathcal{X}_A$ *with* $\mathrm{P}_M(X_A \in \mathcal{A}) > 0$ *such that for* $x_A \in \mathcal{A}$

$$\mathrm{P}_M(X_B \mid \mathrm{do}(X_A = x_A)) \neq \mathrm{P}_M(X_B \mid X_A = x_A).$$

**Remark 2.13.** (1) Note that a common cause has to be an endogenous variable instead of an exogenous variable, because we treat exogenous variables as non-causal (non-intervenable). For example, $X_C$ is a common cause of $X_A$ and $X_B$ according to $M^1$ and $M^2$, but not according to $M^3$ and $M^3$, whose graphs are shown in Figure 5.

(2) If $X_A$ and $X_B$ have confounding bias, then there must be a bidirected edge between $A$ and $B$ in $G(M_{\setminus(V\setminus\{A,B\})})$. The bidirected edge can come from either

[12] "First" and "second" here refer to the order of applying the operations on the SCM, which may not coincide with the chronological order of these operations in the data-generating process that the SCM is modeling.

[13] Providing formal definitions of these concepts for cyclic models is an open research question and is not within the scope of the current manuscript.

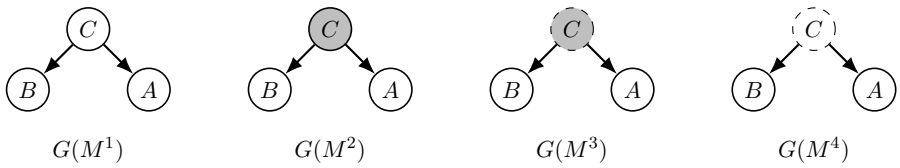



Figure 5: The causal graphs of the SCMs $M^i$, $i = 1, 2, 3, 4$, where $X_C$ is a common cause of $X_A$ and $X_B$ according to $M^1$ and $M^2$ but not to $M^3$ and $M^4$.

(i) a common cause $X_C$, or

(ii) a non-intervenable variable $X_E$ such that there are directed paths from $E$ to $A$ and $B$ in $G(M)^a$, which do not intersect $B$ and $A$ respectively,[14] or

(iii) any combination of the above two items.

If there is no bidirected edge between $A$ and $B$ in $G(M_{\setminus(V\setminus\{A,B\})})$, then by the do-calculus there is no confounding bias between $X_A$ and $X_B$ according to $M$, i.e., we have for a.a. $x_A \in \mathcal{X}_A$ (under the assumption that $B \notin \mathrm{Anc}_{G(M_{\setminus(V\setminus\{A,B\})})}(A)$)

$$\mathrm{P}_M(X_B \mid \mathrm{do}(X_A = x_A)) = \mathrm{P}_M(X_B \mid X_A = x_A).$$

(3) In the potential outcome literature, the unconfoundedness assumption is usually stated as (for the special case of finite discrete outcome variables and binary "treatments") that $X_B$ does not cause $X_A$ (i.e. $X_A(x_B) = X_A$ for all $x_B$) and

$$\forall x_A \in \{0, 1\} : X_A \perp\!\!\!\perp X_B(x_A). \tag{1}$$

If we assume that there is an underlying acyclic SCM $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ inducing the potential outcomes $X_A$ and $X_B(x_A)$, then equation (1) (under a positivity assumption) implies

$$\mathrm{P}_M(X_B \mid \mathrm{do}(X_A = x_A)) = \mathrm{P}_M(X_B \mid X_A = x_A) \text{ for all } x_A \in \{0, 1\}. \tag{2}$$

This means that $X_A$ and $X_B$ have no confounding bias according to $M$ in the sense of Definition 2.12. See Remark B.1 for a proof.

# 3 Theory

In this section, we develop conditioning operations for both SCMs and DMGs. We first introduce s-SCMs, which explicitly encode selection mechanisms, and then define the corresponding conditioning operations and analyze their main properties: the induced causal semantics, closure of relevant model classes, commutation with marginalization, intervention, and further conditioning, the associated loss of information, graphical separation criteria and Markov properties, and the compatibility between the two conditioning operations. After the mathematical development, we conclude with several important caveats on the modeling side.

In the whole section, we make the following assumption (see Remark B.2):

[14] We shall show in the next section that this can represent selection bias. Therefore, a bidirected edge represents the possible existence of "non-causal dependency" between $X_A$ and $X_B$, which can arise from either common cause, selection bias, or any combination of the two, or in other ways.

**Assumption 3.1.** $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ is a simple SCM such that $\mathrm{P}_M(X_S \in \mathcal{S}) > 0$ for some $S \subseteq V$ and measurable subset $\mathcal{S} \subseteq \mathcal{X}_S$.

## 3.1 SCM with selection mechanism

First, we give a definition for SCMs with selection mechanisms.

**Definition 3.2** (SCMs with selection mechanism)**.** *We call $M^{\mathcal{S}} \coloneqq (M, X_S \in \mathcal{S})$ an **s-SCM** or **SCM with a selection mechanism**, where $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ is an SCM, $S \subseteq V$ is a subset of endogenous nodes, and $\mathcal{S} \subseteq \mathcal{X}_S$ is a measurable subset. We call $M^{\mathcal{S}}$ a simple s-SCM if $M$ is simple. The causal semantics of $M^{\mathcal{S}}$ are defined as:*[15]

*(1) Observational distribution:*

$$\mathrm{P}_{M^{\mathcal{S}}}(X_V) \coloneqq \mathrm{P}_M(X_V \mid X_S \in \mathcal{S});$$

*(2) Interventional distributions: for $T \subseteq V \setminus S$ and $x_T \in \mathcal{X}_T$ with $\mathrm{P}_M(X_S \in \mathcal{S} \mid \mathrm{do}(X_T = x_T)) > 0$, we define*

$$\mathrm{P}_{M^{\mathcal{S}}}(X_{V\setminus T} \mid \mathrm{do}(X_T = x_T)) \coloneqq \mathrm{P}_M(X_{V\setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S}).$$

If $\mathrm{P}_M(X_S \in \mathcal{S}) = 1$, then $M^{\mathcal{S}} = (M, X_S \in \mathcal{S})$ is observationally equivalent to $M$. We can draw a causal graph of an SCM with a selection mechanism by Definition A.3, and in addition use triangles to represent the nodes in $S$ like in Figure 2.

To gain some intuition, one can imagine a simple SCM as representing a data-generating process where the $i$-th sample is generated as follows: first sampling $X_w^{(i)} \sim \mathrm{P}(X_w)$ for each $w \in W$, and then using the solution function $g : \mathcal{X}_W \to \mathcal{X}_V$ to generate $X_v^{(i)}$ for each $v \in V$. An SCM with a selection mechanism is adding a rejection step to the above sampling procedure. More precisely, we have the following rejection sampler Algorithm 3.3. This sampler generates the observational distribution of $M^{\mathcal{S}}$. To generate interventional distributions of $M^{\mathcal{S}}$, one just needs to replace $M$ with the corresponding intervened submodel $M_{\mathrm{do}(X_T = x_T)}$, which changes the solution function but leaves the other parts of the algorithm invariant.

## 3.2 Conditioning operation for SCMs

### 3.2.1 Definition

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

(i) merging exogenous variables that become dependent given the observation $X_S \in \mathcal{S}$;

(ii) updating the exogenous probability distribution $\mathrm{P}(X_W)$ to the posterior $\mathrm{P}_M(X_W \mid X_S \in \mathcal{S})$ given the observation $X_S \in \mathcal{S}$;

(iii) marginalizing out the selection variables $X_S$.[16]

[15] We do not specify the counterfactual semantics of s-SCMs.

[16] Selection variables are marginalized out because we consider latent selection.

**Algorithm 3.3** Sampler for an SCM $M$ with a selection mechanism $X_S \in \mathcal{S}$

**Require:** $n \geq 1$
  $i \leftarrow 1$
  **while** $i \leq n$ **do**
    **for** each $w \in W$ **do**
      sample $X_w^{(i)} \sim \mathrm{P}_M\left(X_w\right)$
    **end for**
    **for** each $v \in V$ **do**
      calculate $X_v^{(i)} \leftarrow g_v\left(X_W^{(i)}\right)$
    **end for**
    **if** $X_S^{(i)} \in \mathcal{S}$ **then**
      output $X_V^{(i)}$
      $i \leftarrow i+1$
    **end if**
  **end while**

Before giving a formal definition of the conditioned SCM, we discuss item (i). For the reason why we need to consider merging exogenous random variables, see Section D.2. There is a "finest" partition of $W$ given $X_S \in \mathcal{S}$:

**Lemma 3.4** (Finest partition)**.** *Let $\mathfrak{P}_{\mathcal{S}}$ denote the set of partitions $\mathcal{I} = \{I_1, \ldots, I_p\}$ of $W$ such that $\{X_{I_i}\}_{i=1}^p$ are mutually independent under $\widetilde{\mathrm{P}}(X_W) = \mathrm{P}(X_W \mid X_S \in \mathcal{S})$. Then there exists $\mathcal{H} \in \mathfrak{P}_{\mathcal{S}}$ such that $\mathcal{H}$ is a finer partition than any other partition $\mathcal{I} \in \mathfrak{P}_{\mathcal{S}}$.*

We now present the formal definition of the conditioned SCM.

**Definition 3.5** (Conditioned SCM)**.** *Assume Assumption 3.1. Let $g : \mathcal{X}_W \to \mathcal{X}_V$ and $g^S : \mathcal{X}_{V \setminus S} \times \mathcal{X}_W \to \mathcal{X}_S$ be the (essentially unique) solution functions of $M$ w.r.t. $V$ and $S$ respectively. We define the* ***conditioned SCM*** *$M_{|X_S \in \mathcal{S}} \coloneqq \left(\widehat{V}, \widehat{W}, \widehat{\mathcal{X}}, \widehat{\mathrm{P}}, \widehat{f}\right)$ by:*

*(i)* $\widehat{V} \coloneqq V \setminus S$*;*

*(ii)* $\widehat{W} \coloneqq \{\widehat{w}_1, \ldots, \widehat{w}_n\}$ *is the finest partition of $W$ such that* $\mathrm{P}_M(X_W \mid X_S \in \mathcal{S}) = \bigotimes_{i=1}^n \mathrm{P}_M(X_{\widehat{w}_i} \mid X_S \in \mathcal{S})$*;*

*(iii)* $\widehat{\mathcal{X}} \coloneqq \mathcal{X}_{\widehat{V}} \times \mathcal{X}_{\widehat{W}} \coloneqq \mathcal{X}_{\widehat{V}} \times \bigtimes_{i=1}^n \mathcal{X}_{\widehat{w}_i}$*, where* $\mathcal{X}_{\widehat{w}_i} \coloneqq \bigtimes_{w \in \widehat{w}_i} \mathcal{X}_w$*;*

*(iv)* $\widehat{\mathrm{P}} \coloneqq \bigotimes_{i=1}^n \widehat{\mathrm{P}}(X_{\widehat{w}_i})$*, where* $\widehat{\mathrm{P}}(X_{\widehat{w}_i}) \coloneqq \mathrm{P}_M(X_{\widehat{w}_i} \mid X_S \in \mathcal{S})$*;*

*(v)* $\widehat{f}(x_{\widehat{V}}, x_{\widehat{W}}) \coloneqq f_{\widehat{V}}(x_{\widehat{V}}, g^S(x_{\widehat{V}}, x_{\widehat{w}_1}, \ldots, x_{\widehat{w}_n}), x_{\widehat{w}_1}, \ldots, x_{\widehat{w}_n})$.

It is easy to check that $M_{|X_S \in \mathcal{S}}$ is indeed an SCM. We mark nodes in $\mathrm{Anc}_{G(M)}(S)$ as non-intervenable in $M_{|X_S \in \mathcal{S}}$.

**Remark 3.6.** (1) In Definition 3.5, $M_{|X_S \in \mathcal{S}}$ actually depends on the choice of $g^S$, but different versions are equivalent (in the sense of Definition 2.8). 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 implicitly work with equivalence classes of SCMs. Note that if $M$ and $\widetilde{M}$ are equivalent, then $M_{|X_S \in \mathcal{S}}$ is equivalent to $\widetilde{M}_{|X_S \in \mathcal{S}}$.

(2) The definition of $\widehat{W}$ does not depend on the choice of $g$ but depends on $S$ and $\mathcal{S}$. Note that if $\mathrm{P}_M(X_S \in \mathcal{S}) = 1$, then $\widehat{W} \cong W$ and conditioning on the selection mechanism $X_S \in \mathcal{S}$ reduces to marginalizing out $S$.

(3) If $w \in W \setminus \mathrm{Anc}_{G^a(M)}(S)$ or $w \in W \setminus \mathrm{Anc}_{G^a(M_{\setminus(V\setminus S)})}(S)$, then there exists $\widehat{w}_i$ such that $\widehat{w}_i = \{w\}$. In other words, if node $w$ is not an ancestor of $S$ in $G^a(M)$ or is not an ancestor of $S$ in $G^a(M_{\setminus(V\setminus S)})$, then node $w$ is not merged with any other nodes in $W$. In these cases, one has $\widehat{\mathrm{P}}(X_{\widehat{w}_i}) = \mathrm{P}(X_w)$.

(4) Since marginalization preserves simplicity and P-null sets are also $\mathrm{P}_M(X_W \mid X_S \in \mathcal{S})$-null sets, $M_{|X_S\in\mathcal{S}}$ is simple (cf. Proposition 3.10).

**Notation 3.7.** *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.*

**Example 3.8** (Example 1.1 continued)**.** We consider Example 1.1 in the Introduction. Let $M$ be the SCM in Example 1.1. Then $\widetilde{M} = (M_{|S_0=0})_{\setminus\{B_0,E_0\}} = (M_{\setminus\{B_0,E_0\}})_{|S_0=0}$.

**Example 3.9** (conditioning operation for SCMs)**.** Consider the following SCMs with nonzero real coefficients $a_i$ for $i = 1, \ldots, 6$ such that $a_5 + a_6 \neq 0$:

$$M^1 : \begin{cases} X = E_1 \sim \mathrm{Uni}([0,1]) \\ Y = E_2 \sim \mathrm{Uni}([0,1]) \\ S = X + Y \\ Z_1 = a_1 X + E_3 \\ Z_2 = a_2 Z_1 + E_4 \\ Z_3 = a_3 Z_1 + a_4 Z_2 + a_5 S + a_6 Y + E_5 \end{cases} \qquad M^2 : \begin{cases} X = E_1 \sim \mathrm{Uni}([0,1]) \\ Y = E_2 \sim \mathrm{Uni}([0,1]) \\ S = \mathbb{1}(X + Y \geq 0.8) \\ Z_1 = a_1 X + E_3 \\ Z_2 = a_2 Z_1 + E_4 \\ Z_3 = a_3 Z_1 + a_4 Z_2 + a_5 S + a_6 Y + E_5. \end{cases}$$

With $\mathrm{pr}_i$ denoting the projection to the $i$-th coordinate and $D \coloneqq \{(x,y) \in [0,1]^2 : x+y \geq 0.8\}$, we then have

$$M^1_{|S\geq 0.8} : \begin{cases} E_{1,2} \sim \mathrm{Uni}(D) \\ X = \mathrm{pr}_1(E_{1,2}) \\ Y = \mathrm{pr}_2(E_{1,2}) \\ Z_1 = a_1 X + E_3 \\ Z_2 = a_2 Z_1 + E_4 \\ Z_3 = a_3 Z_1 + a_4 Z_2 + a_5 X \\ \qquad + (a_5 + a_6) Y + E_5, \end{cases} \qquad M^2_{|S=1} : \begin{cases} E_{1,2} \sim \mathrm{Uni}(D) \\ X = \mathrm{pr}_1(E_{1,2}) \\ Y = \mathrm{pr}_2(E_{1,2}) \\ Z_1 = a_1 X + E_3 \\ Z_2 = a_2 Z_1 + E_4 \\ Z_3 = a_3 Z_1 + a_4 Z_2 \\ \qquad + a_5 + a_6 Y + E_5. \end{cases}$$

We draw the (augmented) causal graphs of $M^1$, $M^2$, $M^1_{|S\geq 0.8}$ and $M^2_{|S=1}$ as shown in Figure 6. Note that

(i) bidirected edges can not only represent latent common causes but also latent selection bias;

(ii) given two SCMs with the same causal graphs, the conditioned SCMs can have different graphs.

If one applies Definition 2.2 to perform interventions on ancestors of $S$ in $M^1_{|S\geq 0.8}$ and $M^2_{|S=1}$ (e.g., $X$ and $Y$ in Figure 6), the interventional distribution will correspond to a *counterfactual* distribution of $M$. For instance, $\mathrm{P}_{M^1_{|S\geq 0.8}}(Z_3 \mid \mathrm{do}(X = x)) = \mathrm{P}_{M^1}(Z_3(x) \mid S \geq 0.8) \neq \mathrm{P}_{M^1}(Z_3 \mid \mathrm{do}(X = x), S \geq 0.8)$. See Theorem 3.14 and Section 3.4 for details. This is the reason why we mark the ancestors of $S$ as dashed (non-intervenable) in $G^a(M^1_{|S\geq 0.8})$, $G^a(M^2_{|S=1})$, $G(M^1_{|S\geq 0.8})$ and $G(M^2_{|S=1})$.

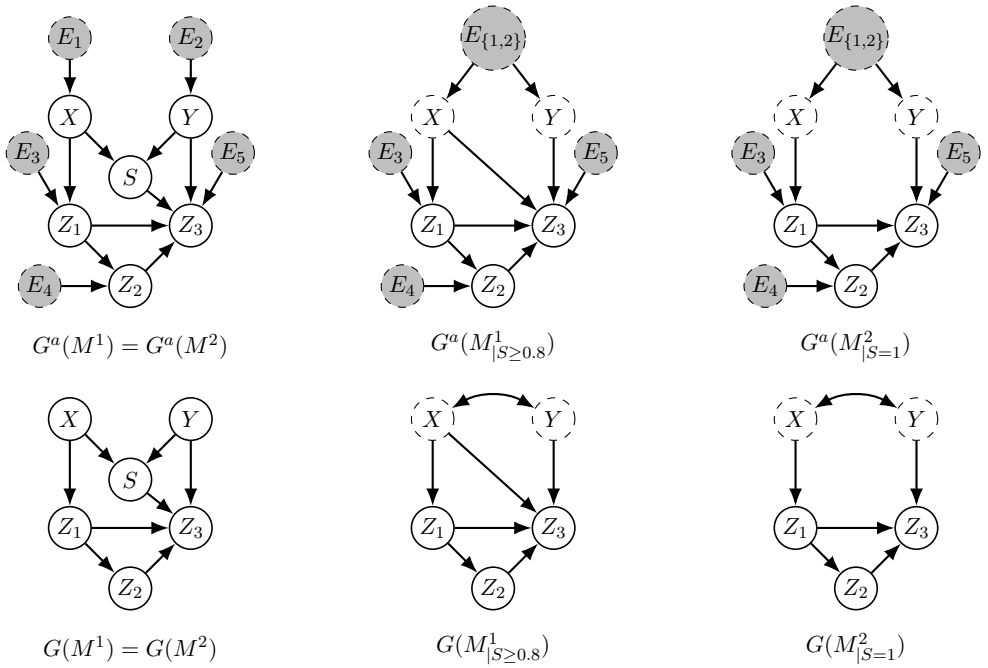



Figure 6: Graphical representation of conditioning on $S \geq 0.8$ and $S = 1$ respectively in $M^1$ and $M^2$. First, merge the exogenous ancestors of $S$, i.e., $E_1$ and $E_2$, to obtain a merged node $E_{\{1,2\}}$. Then update the exogenous probability distribution $\mathrm{P}(E_1, E_2)$ to the posterior $\mathrm{P}_{M^1}(E_1, E_2 \mid S \geq 0.8)$ and $\mathrm{P}_{M^2}(E_1, E_2 \mid S = 1)$. Finally, marginalize out the node $S$. After conditioning, $X$ and $Y$ are dashed, since we mark them as non-intervenable.

### 3.2.2 Properties

We derive some mathematical properties of the conditioning operation. First, we note that the conditioning operation preserves the simplicity, linearity, and acyclicity of SCMs.

**Proposition 3.10** (Simple, acyclic, linear SCMs and conditioning)**.** *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 $M_{|X_S\in\mathcal{S}}$.*

This implies that opting for simple/acyclic/linear SCMs as a model class and performing model abstraction through the conditioning operation will consistently maintain one within the chosen model class. This convenience proves valuable in practical applications, where adherence to the specific model class is often desired.

**Remark 3.11.** Bongers et al. (2021, Proposition 8.2) show that the class of simple SCMs is closed under marginalization, perfect intervention, and the twinning operation. Here we show that the class of simple SCMs is also closed under the conditioning operation of Definition 3.5 (ignoring the subtlety that the ancestors of the conditioning variables became non-intervenable).

In the causal inference community, it is well known that conditioning and intervention do not commute. For pedagogical purposes, we give an example where $T \cap \mathrm{Anc}_{G(M)}(S) \neq \emptyset$ and the intervention and conditioning do not commute. See also, e.g., Mathur and Shpitser (2024) for an example.

**Example 3.12** (Conditioning and intervention do not commute)**.** Consider a linear SCM with $\mathrm{P}(E_T), \mathrm{P}(E_X), \mathrm{P}(E_Y)$ such that $\mathrm{P}_M(X = x) > 0$ and $\mathrm{P}_M(X = x \mid \mathrm{do}(T = t)) > 0$ for $t = 0, 1$ and structural equations

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

In $M$, if we first condition on $X = x$ and second intervene on $T$ (despite $T$ being considered non-intervenable as $T \in \mathrm{Anc}_{G(M)}(X)$), then we have

$$\mathrm{E}_{(M_{|X=x})_{\mathrm{do}(T=1)}}[Y] - \mathrm{E}_{(M_{|X=x})_{\mathrm{do}(T=0)}}[Y] = \alpha + \beta.$$

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

$$\mathrm{E}_{(M_{\mathrm{do}(T=1)})_{|X=x}}[Y] - \mathrm{E}_{(M_{\mathrm{do}(T=0)})_{|X=x}}[Y] = \beta.$$

In general, conditioning and intervention do not commute.

Conditioning does commute with interventions on the non-ancestors of the conditioned variables.

**Lemma 3.13** (Conditioning and intervention)**.** *Assume Assumption 3.1. Let* $T \subseteq V \setminus \mathrm{Anc}_{G(M)}(S)$ *and* $x_T \in \mathcal{X}_T$*. Then we have*

$$\left(M_{\mathrm{do}(X_T = x_T)}\right)_{|X_S \in \mathcal{S}} \equiv \left(M_{|X_S \in \mathcal{S}}\right)_{\mathrm{do}(X_T = x_T)}.$$

Note that in the above lemma, 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 next presented theorem characterizes the causal semantics of conditioned SCMs in terms of the original SCM with selection mechanisms.

**Theorem 3.14** (Main result I: Causal semantics of conditioned SCMs)**.** *Assume Assumption 3.1 and write* $O := V \setminus S$*. Let* $T_i \subseteq O$ *and* $x_{T_i} \in \mathcal{X}_{T_i}$ *for* $i = 1, \ldots, n$*. Then we have*

$$\mathrm{P}_{M_{|X_S \in \mathcal{S}}}\big(\{X_{O \setminus T_i}(x_{T_i})\}_{1 \leq i \leq n}\big) = \mathrm{P}_M\big(\{X_{O \setminus T_i}(x_{T_i})\}_{1 \leq i \leq n} \mid X_S \in \mathcal{S}\big).$$

By noticing that $\mathrm{P}_M(X_{O\setminus T}(x_T)) = \mathrm{P}_M(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T))$ and $\mathrm{P}_M(X_{O\setminus T}(x_T) \mid X_S \in \mathcal{S}) = \mathrm{P}_M(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S})$ provided $T \cap \mathrm{Anc}_{G(M)}(S) = \emptyset$, we have the following result.

**Corollary 3.15.** *Assume Assumption 3.1 and write $O \coloneqq V \setminus S$. Then we have:*

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

*(2)* ***Interventional****: Let $T \subseteq O$ be $T = T_1 \dot\cup T_2$ such that $T_1 \subseteq V \setminus \mathrm{Anc}_{G(M)}(S)$ and $T_2 \subseteq \mathrm{Anc}_{G(M)}(S) \setminus S$. For $x_T \in \mathcal{X}_T$,*

$$\mathrm{P}_{M_{|X_S\in\mathcal{S}}}\left(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T)\right) = \mathrm{P}_M\left(X_{O\setminus T}(x_{T_2}) \mid \mathrm{do}(X_{T_1} = x_{T_1}), X_S \in \mathcal{S}\right).$$

*(3)* ***Counterfactual via twinning****: Let $T \subseteq O$ be $T = T_1 \dot\cup T_2$ such that $T_1 \subseteq V \setminus \mathrm{Anc}_{G(M)}(S)$ and $T_2 \subseteq \mathrm{Anc}_{G(M)}(S) \setminus S$. Let $\widetilde{T} \subseteq V'$ be $\widetilde{T} = T_3 \dot\cup T_4$ such that $T_3 \subseteq (V \setminus \mathrm{Anc}_{G(M)}(S))'$ and $T_4 \subseteq (\mathrm{Anc}_{G(M)}(S) \setminus S)'$.[17] For any $x_T \in \mathcal{X}_T$ and $x_{\widetilde{T}} \in \mathcal{X}_{\widetilde{T}}$,*

$$\begin{aligned}&\mathrm{P}_{\left(M_{|X_S\in\mathcal{S}}\right)^{\mathrm{twin}}}(X_{(O\cup O')\setminus(T\cup\widetilde{T})} \mid \mathrm{do}(X_T = x_T, X_{\widetilde{T}} = x_{\widetilde{T}}))\\ &= \mathrm{P}_{M^{\mathrm{twin}}}(X_{O\setminus T}(x_{T_2}), X_{O'\setminus\widetilde{T}}(x_{T_4}) \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_3} = x_{T_3}), X_S \in \mathcal{S}).\end{aligned}$$

We see that the conditioned SCM faithfully encapsulates the observational distribution of every observed endogenous variable under selection $X_S \in \mathcal{S}$, and (taking $T_2 = T_4 = \emptyset$) the causal semantics of the non-ancestors of $S$ under selection $X_S \in \mathcal{S}$, in accordance with the original SCM. Therefore, the simplified abstracted model yields identical results as the original more intricate model with the selection mechanism $X_S \in \mathcal{S}$ as long as one does not consider $\mathrm{P}_M(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S})$ where $T \cap \mathrm{Anc}_{G(M)}(S) \neq \emptyset$.

One may wonder if it is possible to modify the definition of $M_{|X_S\in\mathcal{S}}$ in such a way that we have, e.g., $\mathrm{P}_{M_{|X_S\in\mathcal{S}}}\left(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T)\right) = \mathrm{P}_M\left(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S}\right)$ also for some $T \subseteq O$ such that $T \cap \mathrm{Anc}_{G(M)}(S) \neq \emptyset$. We show in Section D.1 that it is impossible to find an SCM that preserves the causal semantics of the ancestors of latent selection variables in general.

**Remark 3.16** (Not all data-generating processes with causal interpretation can be modeled by SCMs)**.** Proposition D.1 shows that there exist simple s-SCMs that cannot be modeled by simple SCMs. It suggests that not all data-generating processes with certain causal interpretations can be modeled by SCMs. Such examples also exist in the equilibrium behavior of dynamical systems and functional laws in physics and chemistry (Blom et al., 2020).

In addition, by Theorem 3.14, we have the following simple result. The conditional operation is an operation with a constructive definition that preserves as much causal information as possible.

**Theorem 3.17** (Conditioning operation preserves as much causal information as possible)**.** *There are no mappings $(M, X_S \in \mathcal{S}) \mapsto \widetilde{M}$ that preserve more causal information than*

[17] $V'$ means a copy of $V$. See Definition A.1.

$(M, X_S \in \mathcal{S}) \mapsto M_{|X_S \in \mathcal{S}}$ *in the following sense. Assume that the mapping* $(M, X_S \in \mathcal{S}) \mapsto \widetilde{M}$ *is such that for all* $T_i \subseteq V$ *and* $x_{T_i} \in \mathcal{X}_{T_i}$

$$\mathrm{P}_{\widetilde{M}}\big(\{X_{O\setminus T_i}(x_{T_i})\}_{1\leq i\leq n}\big) = \mathrm{P}_M\big(\{X_{O\setminus T_i}(x_{T_i})\}_{1\leq i\leq n} \mid X_S \in \mathcal{S}\big),$$

*and furthermore for some* $T \subseteq \mathrm{Anc}_{G(M)}(S)$ *and* $x_T \in \mathcal{X}_T$

$$\mathrm{P}_{\widetilde{M}}\left(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T)\right) = \mathrm{P}_M\left(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S}\right).$$

*Then it holds*

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

The subsequent result establishes the commutativity of conditioning and marginalization.

**Proposition 3.18** (Conditioning and marginalization commute)**.** *Assume Assumption 3.1 and let* $L \subseteq V \setminus S$*. Then we have* $(M_{\setminus L})_{|X_S\in\mathcal{S}} \equiv (M_{|X_S\in\mathcal{S}})_{\setminus L}$*.*

Suppose that we have a selection mechanism $X_{S_1\cup S_2} \in \mathcal{S}_1 \times \mathcal{S}_2$, then we can generate three "versions" of the conditioned SCMs:

(i) applying the single conditioning operation w.r.t. $X_{S_1\cup S_2} \in \mathcal{S}_1 \times \mathcal{S}_2$ to get $M_{|\mathcal{S}_1\times\mathcal{S}_2}$;

(ii) first conditioning on $X_{S_1} \in \mathcal{S}_1$ and second conditioning on $X_{S_2} \in \mathcal{S}_2$ to get $(M_{|\mathcal{S}_1})_{|\mathcal{S}_2}$;

(iii) first conditioning on $X_{S_2} \in \mathcal{S}_2$ and second conditioning on $X_{S_1} \in \mathcal{S}_1$ to get $(M_{|\mathcal{S}_2})_{|\mathcal{S}_1}$.

The following proposition demonstrates that these three versions are counterfactually equivalent (and hence empirically indistinguishable).

**Proposition 3.19** (Conditioning and conditioning commute)**.** *Assume Assumption 3.1 with* $S = S_1 \dot\cup S_2$ *and* $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$ *where* $\mathcal{S}_1 \subseteq \mathcal{X}_{S_1}$ *and* $\mathcal{S}_2 \subseteq \mathcal{X}_{S_2}$ *are 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 and induce the same laws of potential outcomes. Also,* $G(M_{|\mathcal{S}_1\times\mathcal{S}_2})$ *is a subgraph of* $G((M_{|\mathcal{S}_1})_{|\mathcal{S}_2})$ *and* $G((M_{|\mathcal{S}_2})_{|\mathcal{S}_1})$*. Furthermore, if*

*(i)* $\mathrm{Anc}_{G^a(M_{\setminus(V\setminus S_1)})}(S_1) \cap \mathrm{Anc}_{G^a(M_{\setminus(V\setminus S_2)})}(S_2) = \emptyset$*, or*

*(ii) we have*

$$\mathrm{P}\left(X_W \in \left(g_{S_1}^{-1}(\mathcal{S}_1) \triangle g_{S_2}^{-1}(\mathcal{S}_2)\right)\right) = 0,$$

*then* $(M_{|\mathcal{S}_1})_{|\mathcal{S}_2} \equiv (M_{|\mathcal{S}_2})_{|\mathcal{S}_1} \equiv M_{|\mathcal{S}_1\times\mathcal{S}_2}$*.*

This result implies that the ordering of applying the conditioning operation does not matter up to counterfactual equivalence of SCMs. Therefore, there is essentially no ambiguity in referring to $M_{|X_S\in\mathcal{S}}$ when the non-empty set $S$ is not a singleton. For marginalization (Bongers et al., 2021, Proposition 5.4), a stronger property holds: marginalizing out variables in different orderings yields equivalent marginal SCMs. Overall, marginalization and conditioning commute both with each other and with themselves up to counterfactual equivalence. So, given a set of latent variables and latent selection mechanisms, irrespective of the intermediate steps taken, one consistently arrives at counterfactually indistinguishable models via marginalization and the conditioning operation. This underscores the robustness and reliability of the overall procedure for model abstraction.

**Example 3.20** (Iterative conditioning and joint conditioning)**.** We give an example showing that applying the conditioning operation iteratively with different orders gives non-equivalent conditioned SCMs and joint conditioning gives a finer model than iterative conditioning does.

Consider an SCM

$$M : \begin{cases} X_{w_1} \sim \mathrm{Uni}\{1,2,3\}, X_{w_2} \sim \mathrm{Uni}\{1,2\}, \\ X_{S_1} = (X_{w_1}, X_{w_2}), X_{S_2} = (X_{w_1}, X_{w_2}). \end{cases}$$

Take $\mathcal{S}_1 \coloneqq \{(2,1),(2,2),(3,1),(3,2)\}$ and $\mathcal{S}_2 \coloneqq \{(1,2),(2,1),(2,2)\}$. Write

$$\begin{aligned} (M_{|\mathcal{S}_1})_{|\mathcal{S}_2} &= (V^{12}, W^{12}, \mathcal{X}^{12}, f^{12}, \mathrm{P}^{12}) \\ (M_{|\mathcal{S}_2})_{|\mathcal{S}_1} &= (V^{21}, W^{21}, \mathcal{X}^{21}, f^{21}, \mathrm{P}^{21}) \\ M_{|\mathcal{S}_1 \times \mathcal{S}_2} &= (V^{1\times 2}, W^{1\times 2}, \mathcal{X}^{1\times 2}, f^{1\times 2}, \mathrm{P}^{1\times 2}). \end{aligned}$$

Then we have $W^{12} = \{w_1, w_2\} = W$, $W^{21} = \{\{w_1, w_2\}\}$ (nodes $w_1$ and $w_2$ merge) and $\widehat{W}^{1\times 2} = \{w_1, w_2\} = W$. Therefore, $(M_{|\mathcal{S}_1})_{|\mathcal{S}_2} \not\equiv (M_{|\mathcal{S}_2})_{|\mathcal{S}_1}$ and $(M_{|\mathcal{S}_2})_{|\mathcal{S}_1} \not\equiv M_{|\mathcal{S}_1\times\mathcal{S}_2}$, but $(M_{|\mathcal{S}_1})_{|\mathcal{S}_2} \equiv M_{|\mathcal{S}_1\times\mathcal{S}_2}$. In addition, $M_{|\mathcal{S}_1\times\mathcal{S}_2}$ is finer than $(M_{|\mathcal{S}_2})_{|\mathcal{S}_1}$.

**Remark 3.21.** The phenomenon in Example 3.20 occurs since: two exogenous random variables ($X_{w_1}$ and $X_{w_2}$) merged due to conditioning on $X_{S_2} \in \mathcal{S}_2$ may become independent after further conditioning on $X_{S_1} \in \mathcal{S}_1$, while the conditioned SCM $M_{|X_{S_2}\in\mathcal{S}_2}$ only records the information of the label set $\widehat{W}$ where $w_1$ and $w_2$ become merged but forgets the original label set $W$ where $w_1$ and $w_2$ are distinct. Therefore, $w_1$ and $w_2$ are distinct in $M_{|\mathcal{S}_1\times\mathcal{S}_2}$ but merge in $(M_{|\mathcal{S}_2})_{|\mathcal{S}_1}$.

This problem can be fixed by modifying the definition of SCMs. For example, one can adapt the definition by equipping it with the information of a specific partition of $W$. We do not pursue this approach in the current manuscript and opt in Definition 3.5 for the "joint" version of the conditioning operation.

Often, the selection nodes do not have children, as depicted in Figure 7. Even when the selection nodes have children, one can always create a copy of it, reducing it to the situation where the selection nodes lack children. For example, consider $G$ depicted in Figure 8, where selections occur on $A$ and $C$ by conditioning on a common effect $S_1$ and on $D$ by conditioning on $S_2$. Introducing a copy $\widetilde{S}$ of $(S_1, S_2)$, or setting $\widetilde{S} = \mathbb{1}_{\{(S_1,S_2)\in\mathcal{S}\}}$, yields the causal graph $\widetilde{G}$, where conditioning is performed on $\widetilde{S}$ instead of $(S_1, S_2)$. Consequently, in the causal graph $\widetilde{G}$, the selection node $\widetilde{S}$ does not have children. The following lemma validates this construction.

**Lemma 3.22** (Conditioning on binary variable without children)**.** *Let $(M, X_S \in \mathcal{S})$ be a simple s-SCM and $(\widetilde{M}, X_{\widetilde{S}} = 1)$ be another simple s-SCM where $\widetilde{M} \coloneqq (\widetilde{V}, W, \widetilde{X}, \mathrm{P}, \widetilde{f})$ is such that $\widetilde{V} = V \dot\cup \{\widetilde{S}\}$, $\widetilde{\mathcal{X}} = \mathcal{X} \times \mathcal{X}_{\widetilde{S}} \coloneqq \mathcal{X} \times \{0,1\}$ and $\widetilde{f}_{\widetilde{V}\setminus\widetilde{S}}(x_{\widetilde{V}}, x_W) = f_V(x_V, x_W)$ and $\widetilde{f}_{\widetilde{S}}(x_{\widetilde{V}}, x_W) = \mathbb{1}_{\mathcal{S}}(x_S)$. Then $M_{|X_S\in\mathcal{S}} \equiv (\widetilde{M}_{|X_{\widetilde{S}}=1})_{\setminus S}$.*

Based on the above observation, we give a conditioning operation for if we want to observe the selection variable. Let $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ be a simple SCM. Suppose that we want to condition on $X_S \in \mathcal{S}$ for $S \subseteq V$ and $\mathcal{S} \subseteq \mathcal{X}_S$ so that $\mathrm{P}_M(X_S \in \mathcal{S}) > 0$ but we still want to observe the values of $X_S$. Then a solution is that we introduce a selection variable $X_{\widetilde{S}}$ such that $X_{\widetilde{S}} = \mathbb{1}_{\mathcal{S}}(X_S)$ as we did in Lemma 3.22 and condition on $X_{\widetilde{S}} = 1$.

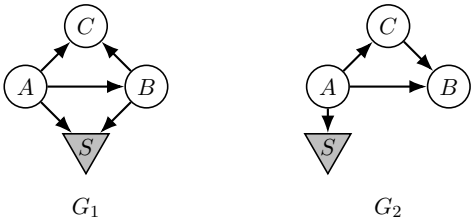



Figure 7: Causal graphs representing selection on node $S$.

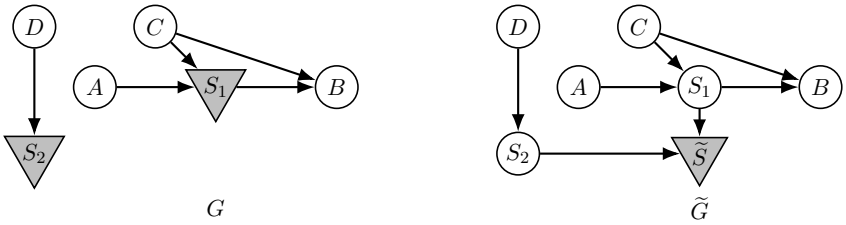



Figure 8: Causal graphs representing selection on nodes $S_1$ and $S_2$, and on node $\widetilde{S}$, respectively.

## 3.3 Conditioning operation for DMGs

The conditioning operation (Definition 3.5) is defined on simple SCMs. For causal modeling purposes, people often use causal graphs to communicate causal knowledge without referring to the precise underlying SCMs. To support this, we give a purely graphical conditioning operation defined on directed mixed graphs (DMGs). The idea is: (i) to add bidirected edges to node pairs that are ancestors of the conditioned nodes or siblings of ancestors of the conditioned nodes, and then (ii) to graphically marginalize the conditioned nodes out.

**Definition 3.23** (Conditioned DMG)**.** *Let $G = (V, E^d, E^b)$ be a DMG consisting of nodes $V$, directed edges $E^d$ and bidirected edges $E^b$. For $S \subseteq V$, we define the conditioned DMG $G_{|S}$ by*

*(1) adding bidirected edges to $G$: $\{a \dashleftarrow\dashrightarrow b : a, b \in \mathrm{Anc}_G(S) \cup \mathrm{Sib}_G(\mathrm{Anc}_G(S))\}$.*[18]

*(2) marginalizing out $S$ and marking ancestors of $S$ as dashed.*

The definition is inspired by the conditioning operation for SCMs. As we shall show, the purely graphical conditioning operation is compatible with the SCM conditioning operation.

**Example 3.24** (DMGs conditioning)**.** We show an example of the purely graphical conditioning operation. Assume that we are given a graph $G$ as shown in Figure 9. Then conditioning on node $V_5$ gives the graph $G_{|V_5}$ shown in Figure 9.

The following result shows that the conditioning graph $G_{|S}$ represents $\sigma$-separations (also $d$-separations) encoded in the original graph $G$ soundly (but not completely in general). The notion of $\sigma$-separation (Definition A.9) is defined for DMGs and reduces to normal $d$-separation (or $m$-separation) of acyclic DMGs (Richardson, 2003) if there are no cycles (Forré

[18] $\mathrm{Sib}_G(v) \coloneqq \{w \in G \mid v \dashleftarrow\dashrightarrow w \text{ is in } G\}$.

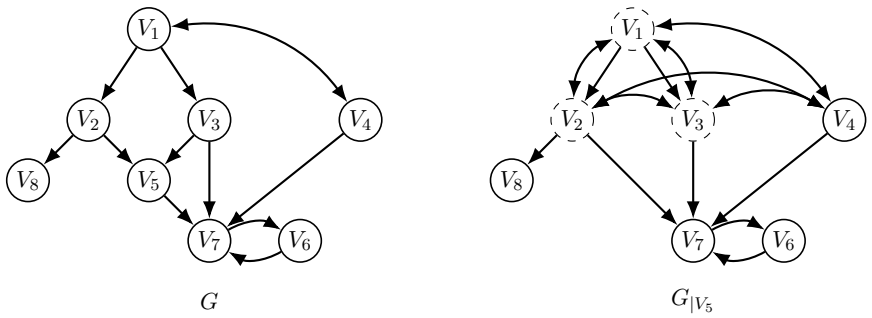



Figure 9: DMG $G$ and its conditioned DMG $G_{|V_5}$ in Example 3.24.

and Mooij, 2017). Note that when there are cycles, $\sigma$-separation implies $d$-separation but not the other direction.

**Theorem 3.25** (Main result II: Graphical separation in conditioning graph)**.** *Let $G = (V, E^d, E^b)$ be a DMG and $S \subseteq V$ a set of nodes. Then for any subsets of nodes $A, B, C \subseteq V$ such that*

$$S \cap (A \cup B \cup C) = \emptyset,$$

*it holds that*

$$A \overset{\sigma/d}{\underset{G_{|S}}{\perp}} B \mid C \quad \Longrightarrow \quad A \overset{\sigma/d}{\underset{G}{\perp}} B \mid C \cup S.$$

*If furthermore $S$ is a singleton set with $\mathrm{Ch}_G(S) = \emptyset$ and $C \cap \mathrm{Anc}_G(S) = \emptyset$, then we have*

$$A \overset{\sigma/d}{\underset{G}{\perp}} B \mid C \cup S \quad \Longrightarrow \quad A \overset{\sigma/d}{\underset{G_{|S}}{\perp}} B \mid C.$$

**Remark 3.26.** If we only consider a singleton conditioning event $\mathcal{S} = \{x_S\}$ (see Example B.6 and Lemma 3.30 for the subtle difference), then we can define another version of the graphical conditioning operation transforming $G$ to $G_{\mathrm{cd}(S)}$ ("cd" represents condition on). It is defined by $G_{\mathrm{cd}(S)} := (G_{\underline{S}})_{|S}$, i.e., we first delete all the arrows emerging from $S$ and then apply the original conditioning operation to $G_{\underline{S}}$. It is easy to see that this new construction can strengthen the above result by removing the assumption that $\mathrm{Ch}_G(S) = \emptyset$. This is similar in spirit to marking constant variables and deterministic dependencies distinctly in a causal graph to capture more conditional independence information escaping from the usual $d$-separation Markov property (see, e.g., Geiger et al. (1990); Spirtes et al. (2001)).

The proof of the second claim in Theorem 3.25 relies on the three assumptions. See Example B.5 in Section B.

**Proposition 3.27** (Graph conditioning commutes with marginalization, conditioning and intervention)**.** *Let $G = (V, E^d, E^b)$ be a DMG.*

*(1) Let $L \subseteq V$ and $S \subseteq V$ be two disjoint subsets of nodes from $G$. Then we have*

$$\left(G_{\backslash L}\right)_{|S} = \left(G_{|S}\right)_{\backslash L}.$$

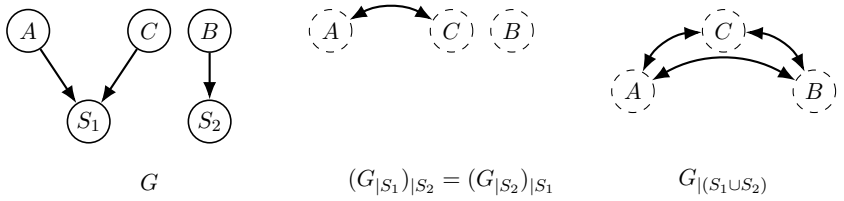
*(2) Let $S_1, S_2 \subseteq V$ be two disjoint subsets. Then we have*

$$\left(G_{|S_1}\right)_{|S_2} = \left(G_{|S_2}\right)_{|S_1} \subseteq G_{|S_1 \cup S_2}.$$

*(3) Let $T \subseteq V$ and $S \subseteq V$ be two disjoint subsets of nodes from $G$ such that $T \cap \mathrm{Anc}_G(S) = \emptyset$. Then we have*

$$\left(G_{\mathrm{do}(T)}\right)_{|S} = \left(G_{|S}\right)_{\mathrm{do}(T)}.$$

**Remark 3.28.** One should note that $\left(G_{|S_1}\right)_{|S_2}$ could be a strict subgraph of $G_{|(S_1 \cup S_2)}$. An example is shown in Figure 10. When knowing that the conditioned set can be decomposed as a Cartesian product, one should use the iterative conditioned graph, which is finer than the jointly conditioned graph. See Section 3.3.1.




$G$ $\quad$ $(G_{|S_1})_{|S_2} = (G_{|S_2})_{|S_1}$ $\quad$ $G_{|(S_1 \cup S_2)}$

Figure 10: $\left(G_{|S_1}\right)_{|S_2} = \left(G_{|S_2}\right)_{|S_1} \subsetneq G_{|(S_1 \cup S_2)}$

The following proposition states that the purely graphical conditioning operation is compatible with the SCM conditioning operation. See Remark B.3 in Section B for some remarks on Proposition 3.29.

**Proposition 3.29** (Main result III: DMG conditioning is compatible with SCM conditioning)**.** *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 $G(M)_{|S}$. If furthermore $S = \{s_1, \ldots, s_n\}$ and $\mathcal{S} = \bigtimes_{i=1}^n \mathcal{S}_i$ with $\mathcal{S}_i \subseteq \mathcal{X}_{s_i}$ measurable for $i = 1, \ldots, n$, then $G(M_{|X_S \in \mathcal{S}})$ is a subgraph of $((G(M)_{|s_1})_{\ldots})_{|s_n}$.*

(Generalized) Directed global Markov properties connect the causal graph $G(M)$ and the induced distribution $\mathrm{P}_M(X_V)$ of the SCM $M$ in the sense that they enable one to read off conditional independence relations from the graph via the $d$-separation (resp. $\sigma$-separation) criterion (Definitions A.7 and A.10). Obviously, $\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_O)$ satisfies the (generalized) directed Markov property relative to $G(M_{|X_S \in \mathcal{S}})$ (Theorems A.8 and A.11). Therefore, the above proposition immediately implies that $\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_O)$ satisfies the Markov property relative to $G(M)_{|S}$ (as Corollary 3.31 states), illustrating the role of the conditioned graph $G(M)_{|S}$ as an effective graphical abstraction. However, one should be aware of the subtlety that one *cannot* directly conclude

$$A \overset{\sigma/d}{\underset{G(M)}{\perp}} B \mid C, S \quad \Longrightarrow \quad X_A \underset{\mathrm{P}_M(X_V)}{\perp\!\!\!\perp} X_B \mid X_C, X_S \in \mathcal{S}$$

even if $\mathrm{P}_M(X_V)$ satisfies the Markov property relative to $G(M)$. See Example B.6 for details.

The following lemma establishes a connection between conditional independence given a variable and a family of conditional independencies given certain events.

**Lemma 3.30.** *Let $X_A, X_B, X_C$ and $X_S$ be random variables defined on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and $X_S$ take values in a standard measurable space $(\mathcal{X}_S, \mathcal{B}_{\mathcal{X}_S})$. Then the first statement implies the second statement:*

*(1) $X_A \perp\!\!\!\perp X_B \mid X_C, X_S \in \mathcal{H}$ for all $\mathcal{H} \in \mathcal{B}_{\mathcal{X}_S}$ with positive probability.*

*(2) $X_A \perp\!\!\!\perp X_B \mid X_C, X_S$.*

This lemma incorporates some classical cases as special instances. For example, if $X_S$ has a countable support $\mathcal{X}_S$, then $X_A \perp\!\!\!\perp X_B \mid X_S = s$ for all $s \in \mathcal{X}_S$ implies $X_A \perp\!\!\!\perp X_B \mid X_S$ (the converse also holds). The converse of the above lemma does not hold as Example B.6 shows. Also given one single $\mathcal{S} \subseteq \mathcal{X}_S$ such that $\mathrm{P}(X_S \in \mathcal{S}) > 0$ and $X_A \perp\!\!\!\perp X_B \mid X_C, X_S \in \mathcal{S}$, we cannot infer $X_A \perp\!\!\!\perp X_B \mid X_C, X_S$ in general.

Understanding these subtleties in conditional independence allows us to better appreciate the following Markov property, which is an easy corollary from Propositions 3.10 and 3.29 and Theorems A.8 and A.11.

**Corollary 3.31.** *Let $M$ be a simple SCM with conditioned SCM $M_{|X_S \in \mathcal{S}}$. Then the uniquely induced distribution $\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_O)$ satisfies the generalized directed global Markov property relative to $G(M)_{|S}$, i.e., for $A, B, C \subseteq O$ we have*

$$A \overset{\sigma}{\underset{G(M)_{|S}}{\perp}} B \mid C \quad \Longrightarrow \quad X_A \underset{\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_O)}{\perp\!\!\!\perp} X_B \mid X_C.^{19}$$

*Furthermore, assume one of the following conditions: (i) $M$ is acyclic; (ii) all endogenous state spaces $\mathcal{X}_v$ are discrete; (iii) $M_{|\mathcal{S}}$ satisfies the third assumption in Theorem A.8. Then $\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_O)$ satisfies the directed global Markov property relative to $G(M)_{|S}$, i.e.,*

$$A \overset{d}{\underset{G(M)_{|S}}{\perp}} B \mid C \quad \Longrightarrow \quad X_A \underset{\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_O)}{\perp\!\!\!\perp} X_B \mid X_C.$$

We can apply the intervention operation to $M_{|X_S \in \mathcal{S}}$ and $G(M)_{|S}$, respectively. Recall that a simple SCM $M$ always satisfies the generalized Markov property w.r.t. $G(M)$. Therefore, we obtain that $(M_{|X_S \in \mathcal{S}})_{\mathrm{do}(X_T = x_T)}$ satisfies the generalized Markov property w.r.t. $(G(M)_{|S})_{\mathrm{do}(T)}$ for any $T \subseteq O$.

The converse of (generalized) directed global Markov properties is $d$-faithfulness (resp. $\sigma$-faithfulness) (Spirtes et al., 2001; Pearl, 2009; Forré and Mooij, 2018), which plays an important role in constraint-based causal discovery algorithms (Spirtes et al., 1995, 1999; Mooij and Claassen, 2020). A natural question arises: how does faithfulness interact with the conditioning operation? Recall that marginalization preserves faithfulness. If $\mathrm{P}_M(X_V)$ is faithful to $G(M)$, then $\mathrm{P}_{M_{\setminus L}}(X_{V \setminus L})$ is faithful to $G(M)_{\setminus L}$ for any $L \subseteq V$. However, this does not generally hold for conditioning. Even if $\mathrm{P}_M(X_V)$ is faithful to $G(M)$, the distribution $\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_{V \setminus S})$ may not be faithful to $G(M)_{|S}$. See Example 3.32 for a simple example.

[19] Recall that $X_A \underset{\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_O)}{\perp\!\!\!\perp} X_B \mid X_C \iff X_A \underset{\mathrm{P}_M(X_V)}{\perp\!\!\!\perp} X_B \mid X_C, X_S \in \mathcal{S}$.

**Example 3.32** (Conditioning does not preserve faithfulness)**.** Consider an SCM $M$ and its conditioned SCM $M_{|X_s=1}$

$$M: \begin{cases} X_{w_1}, X_{w_2} \sim \mathrm{Ber}(0.5), \\ X_a = X_{w_1}, X_s = X_{w_2} \\ X_b = \mathbb{1}(X_{w_2} = 0)X_a, \end{cases} \qquad M_{|X_s=1}: \begin{cases} X_{w_1} \sim \mathrm{Ber}(0.5), X_{w_2} = 1, \\ X_a = X_{w_1}, \\ X_b = \mathbb{1}(X_{w_2} = 0)X_a. \end{cases}$$

It is easy to see that $X_a \underset{\mathrm{P}_{M_{|X_s=1}}(X_{V\setminus\{s\}})}{\perp\!\!\!\perp} X_b$. It holds that $\mathrm{P}_{M_{|X_s=1}}(X_{V\setminus\{s\}})$ is faithful to $G(M_{|X_s=1})$ but not to $G(M)_{|\{s\}}$.

Faithfulness is usually stated in terms of conditional independence given variables while the conditioned SCMs encode conditional independence information given selection variables taking values in some sets. Example B.6 in Section B shows the subtle difference between the two. However, Lemma 3.30 builds the connection between them and allows us to state faithfulness in terms of conditional independence given events.

Putting all the above results in Section 3 together gives us an answer to Questions Q2 and Q3 in the Introduction. For the interaction between SCMs and causal graphs, assume that we have a simple SCM $M$ and only a subset $O$ of $V$ is observable. Denote $V \setminus O$ by $L \dot\cup S$ where $L$ denotes the latent part that is marginalized out and $S$ denotes the latent selection nodes. Fix a measurable set $\mathcal{S} \subseteq \mathcal{X}_S$ with $\mathrm{P}_M(X_S \in \mathcal{S}) > 0$. Define the observable marginalized conditioned SCM

$$M_{O|\mathcal{S}} \coloneqq (M_{\setminus L})_{|\mathcal{S}} \equiv (M_{|\mathcal{S}})_{\setminus L},$$

and observable marginalized conditioned graph

$$G^{[O]} \coloneqq ((G(M))_{\setminus L})_{|S} = ((G(M))_{|S})_{\setminus L},$$

Then we have Figure 11 (dashed arrows mean that the implications are only true under some extra conditions, and the numbers near the arrows correspond to theorems, lemmas, corollaries, and examples) for $A, B, C \subseteq O$.

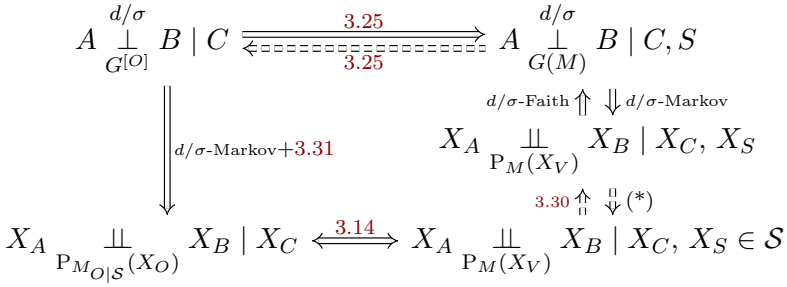



Figure 11: Diagram relating graphical separation and stochastic independence under marginalization and conditioning for a simple SCM $M$.

(*): Example B.6 tells us that this implication does not hold in general, but if $\mathcal{S} = \{x_s\}$ is a singleton set (recall that we assume $\mathrm{P}_M(X_S \in \mathcal{S}) > 0$) then this implication holds.

The diagram shown in Figure 11 still holds if we replace $G^{[O]}$ and $M_{O|\mathcal{S}}$ with $(G^{[O]})_{\mathrm{do}(T)}$ and $(M_{O|S})_{\mathrm{do}(X_T=x_T)}$ respectively.[20]

**Remark 3.33.** To compare observational distributions and interventional distributions in one single graph and therefore derive the general measure-theoretic causal calculus rigorously, one may need the so-called exogenous input variables (or non-stochastic regime indicator variables according to Dawid (2021)) and transitional probability theory (Forré, 2021; Forré and Mooij, 2020). We discuss a conditioning operation for this more general class of models in Appendix D.3. With this, we can generalize measure-theoretic causal calculus and other identification results to the case with latent selection bias via the conditioning operation (cf. Definition D.9), and develop a commutative diagram similar to the one shown above but with exogenous non-stochastic input variables. Since causal information can be alternatively characterized by conditional independence involving regime indicators (Dawid, 2002, 2021), the corresponding diagram for SCMs with input nodes gives us a clearer picture of what causal information can be preserved during the process of model abstraction via the conditioning operation.

### 3.3.1 Conditioning operation for DMGs: explicit modularity and locality

As mentioned in Remark 3.28, when knowing that the conditioning set $\mathcal{S} \subseteq \mathcal{X}_S$ can be decomposed as a Cartesian product $\mathcal{S} = \bigtimes_{i=1}^n \mathcal{S}_i$ with $\mathcal{S}_i \subseteq \mathcal{X}_{s_i}$ and $S = \{s_1, \dots, s_n\}$, we can obtain a finer conditioned graph by iterative conditioning than by joint conditioning. We give a formal definition.

**Definition 3.34** (Conditioned DMG: special case)**.** *Let $G$ be a DMG. For $S = \{s_1, \dots, s_n\} \subseteq V$, we define the conditioned DMG $G_{|\boxtimes S}$ by $((G_{|s_1})_{\dots})_{|s_n}$.*

This definition is more in line with the principle that the SCM expresses the modular structure of causal mechanisms and selection mechanisms. This is particularly relevant when modeling physical systems where the locality principle of special relativity should be respected (both for causal mechanisms and selection mechanisms). Note that the definition of $G_{|\boxtimes S}$ does not depend on the ordering of the iterative conditioning by Proposition 3.27. Theorem 3.25 and Proposition 3.27 all hold if we replace $(\cdot)_{|S}$ with $(\cdot)_{|\boxtimes S}$. By Proposition 3.29, we know that $G(M_{|X_S \in \mathcal{S}})$ is a subgraph of $G(M)_{|\boxtimes S}$ where $S = \{s_1, \dots, s_n\}$ and $\mathcal{S} = \bigtimes_{i=1}^n \mathcal{S}_i$. Furthermore, we have $(G_{|\boxtimes S_1})_{|\boxtimes S_2} = (G_{|\boxtimes S_2})_{|\boxtimes S_1} = G_{|\boxtimes (S_1 \cup S_2)}$. If we introduce a common child $S^*$ of $s_1, \dots, s_n$ and call the extended graph $G^*$, then $G_{|S} = (G^*)_{|S^*} = (G^*)_{|\boxtimes S^*} \supseteq G_{|\boxtimes S}$.

## 3.4 Caveats on modeling interpretation

In the previous subsections, we presented the SCM conditioning operation and DMG conditioning as purely mathematical operations and derived some mathematical properties of them. In this subsection, we 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

[20] One should be careful with the causal interpretation when $T \nsubseteq O \setminus \mathrm{Anc}_{G(M)}(S)$. See Section 3.4 for more details.

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 a 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 perform an intervention (say $\mathrm{do}(T = t)$) on this *fixed selected subpopulation* instead of on the total population.[21] Mathematically, we have

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

where we used the language of potential outcomes and the twinning operation. 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. See also Pearl (2015) for an illustration.

One can think of at least three possible ways to use the conditioning operation for modeling:

(i) marginalize out all the ancestors of the selection nodes, or only consider cases where selection happens on exogenous random variables, so that there is no chance of being tempted to intervene on the ancestors of the selection nodes;

(ii) specify in the conditioned SCM or in its graph which variables are ancestors of the selection nodes in the original SCM and do not apply interventions on them (which is what we opt for in this work);

(iii) one can ignore the issue if one does not mind mixing up the rung-two quantities and rung-three quantities for her tasks, at the risk of introducing confusion about the causal interpretation of the conditioned SCM (which is not recommended).

See Remark B.4 in Section B for some further remarks on modeling interpretation.

# 4 Applications

In this section, we illustrate several applications of the conditioning operation. The conditioning operation has a wide range of uses: all classical results for SCMs, such as identification results (back-door adjustment, do-calculus), apply directly to the conditioned SCMs $M_{|X_S \in \mathcal{S}}$. Using the properties of the conditioning operation, these conclusions for $M_{|X_S \in \mathcal{S}}$ can then be

[21] This is often impossible to do in the real world where time travel is not an option, except if we can "redo" interventions while the exogenous variables remain invariant. Therefore, we prefer not to degrade from rung-2 causal queries to rung-3 ones.

translated back to $(M, X_S \in \mathcal{S})$. In combination with marginalization, conditioning operation also provides a way to interpret a DMG as a causal graph that compactly encodes causal assumptions, with latent details of both latent common causes and latent selection abstracted away.

The examples in this section form a cohesive sequence. It navigates us from the philosophical implications of conditioning (a "generalized Reichenbach's principle"), to the versatility of applying classical results to conditioned SCMs (back-door criterion, ID-algorithm, causal discovery, instrumental variables, mediation analysis), and finally to a concrete practical application of conditioned SCMs to modeling real-world problems (the COVID example).

## 4.1 Reichenbach's principle under latent selection

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). Note that this conclusion holds only when latent selection bias is ruled out, an assumption that is often left implicit.

**Example 4.1** (Reichenbach's principle)**.** Using the conditioning operation, we can generalize and prove the principle under the framework of SCMs in the following way. Assume that $M$ is a simple SCM that has two observed endogenous variables $X$ and $Y$. By the Markov property (Theorem A.11), 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 are subject to latent selection (or any combination of these four possibilities).

**Remark 4.2.** This provides one possible explanation for some real-world scenario in which one can exclude the possibilities of causal effects and common causes between two variables but can still observe the stochastic dependency between them.

## 4.2 Causal identification under latent selection

**Example 4.3** (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 12. Under some assumptions, Pearl's Back-Door Theorem (Pearl, 2009) gives, for $i = 1, 2$, the identification result:[22]

$$\mathrm{P}_{M^i}(Y \mid \mathrm{do}(T = t)) = \int \mathrm{P}_{M^i}(Y \mid X = x, T = t)\mathrm{P}_{M^i}(X \in \mathrm{d}x). \tag{3}$$

Thanks to marginalization and the conditioning operation, we can see $M^1$ and $M^2$ as abstractions of other SCMs, i.e., $M^i = (\widetilde{M}^i_{\setminus L^i})_{|\mathcal{S}^i}$, for SCMs $\widetilde{M}^i$, 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

[22] For simplicity, here we ignore the measure-theoretic subtlety. Indeed, we need to assume $\mathrm{P}_{M^i}(X) \otimes \mathrm{P}_{M^i}(T) \ll \mathrm{P}_{M^i}(X, T)$ and then the identity holds $\mathrm{P}_{M^i}(T)$-a.s. See Forré and Mooij (2025) for more details.

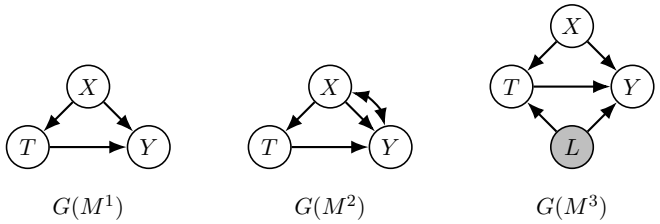



Figure 12: Causal graphs of SCMs $M^1$ and $M^2$ in Example 4.3 and of $M^3$ in Remark 4.5.

sets $\mathcal{S}^i$ with $i = 1, 2$. For both $M^1$ and $M^2$, we present two examples $\widetilde{M}^i_{(j)}$ for $j = 1, 2$, respectively, out of the infinite possibilities in Figure 13.

With the help of Theorem 3.14, we can write (3) as

$$\mathrm{P}_{\widetilde{M}^i}(Y \mid \mathrm{do}(T = t), S^i \in \mathcal{S}^i) = \int \mathrm{P}_{\widetilde{M}^i}(Y \mid X = x, T = t, S^i \in \mathcal{S}^i)\mathrm{P}_{\widetilde{M}^i}(X \in \mathrm{d}x \mid S^i \in \mathcal{S}^i). \tag{4}$$

Thus, the back-door theorem can be applied directly to the conditioned SCM, which is useful especially if the specific latent structure of the SCM is *unknown*.

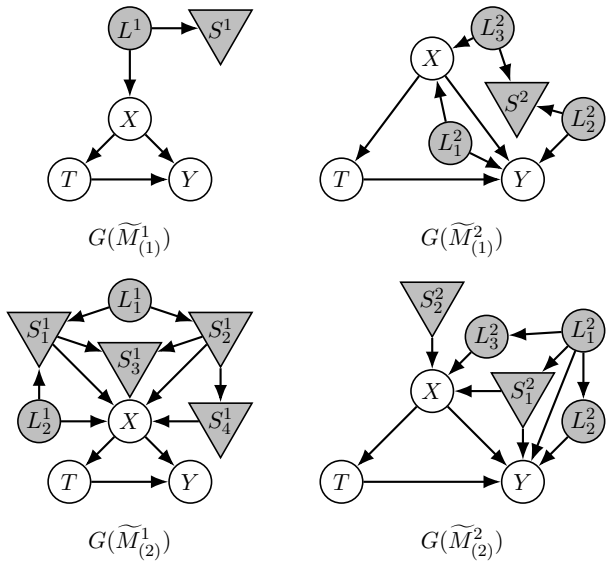



Figure 13: Some possible causal graphs of SCMs $\widetilde{M}^i$ in Example 4.3.

**Remark 4.4.** One can generalize other identification results similarly.

**Remark 4.5.** The result in Example 4.3 differs from that in Correa and Bareinboim (2017), where it is assumed that the explicit causal structure of the selection mechanism is known, allowing identification of the causal effect in the whole population from the selected data. It is a slight generalization of the conditional back-door adjustment in Pearl (2009), which is expressed as $\mathrm{P}(Y = y \mid \mathrm{do}(T = t), S = s) = \sum_x \mathrm{P}(Y = y \mid X = x, T = t, S = s)\mathrm{P}(X = x \mid S = s)$ when certain graphical criteria are met. One difference is that in (4), $\mathcal{S}^i$ may not be a singleton but a general set. The generalized back-door criterion for MAGs cannot be applied

here, since it rules out selection bias explicitly (Maathuis and Colombo, 2015). Note that even if we rule out selection bias, interpreting the graph $G(M^1)$ as a MAG will have different consequences than interpreting it as a causal graph of an SCM. Indeed, $\mathrm{P}(Y \mid \mathrm{do}(T))$ is not identifiable in $G(M^3)$ in Figure 12 while the MAG representation of $G(M^3)$ is syntactically equal to the graph $G(M^1)$.

Pearl's do-calculus is proved to be sound and complete (under some conditions) for identifying interventional distributions in terms of the observational distribution given a causal graph (Pearl, 1995a; Huang and Valtorta, 2006). Using a causal graph and observational distribution as inputs, the ID-algorithm, as a sound and complete algorithm, systematically expresses the target interventional distribution in terms of a functional of the observational distribution, if the target is identifiable and outputs FAIL if not (Tian and Pearl, 2002; Shpitser and Pearl, 2006b; Huang and Valtorta, 2008). Various variants of the ID-algorithm exist, each with different targets and inputs (see e.g., Kivva et al. (2023) and the references therein).

**Example 4.6** (ID-algorithm)**.** One such variant, the s-ID-algorithm, is a sound and complete algorithm for the *s-identification problem*, whose goal is to identify interventional distributions on a subpopulation $(\mathrm{P}(X_A \mid \mathrm{do}(X_T = x_T), X_S = 1))$ given a causal graph with selection mechanism $(G)$ and selected observational distribution $(\mathrm{P}(X_V \mid X_S = 1))$ (Abouei et al., 2024a, Theorem 1, Corollary 2).[23] As we shall see, the conditioning operation can help simplify a part of the original proof (Abouei et al., 2024b, Lemma 5).

Consider the single-variable case $T = \{t\}$. In the setting of Abouei et al. (2024a), there are no latent variables. Therefore, if $T \cap \mathrm{Anc}_G(S) = \emptyset$, then there are no bidirected edges connecting to $t$ in $G_{|S}$, which implies that $\mathrm{P}(X_A \mid \mathrm{do}(X_T = x_T), X_S = 1)$ is identifiable by Tian and Pearl (2002, Theorem 1). Now, assuming that $A \overset{d}{\underset{G_{\underline{T}}}{\perp}} T \mid S$, the second rule of Pearl's do-calculus provides the identification result. Combining these two gives a sound and complete algorithm for the s-identification problem (Abouei et al., 2024a, Theorem 1). The soundness of this algorithm immediately generalizes to settings with latent variables.

Besides, if $T \cap \mathrm{Anc}_G(S) = \emptyset$, then one can also consider identifying the conditional causal effect on the subpopulation $\mathrm{P}(X_A \mid \mathrm{do}(X_T = x_T), X_B, X_S = 1)$ from a graph with latent variables and selected observational distribution $\mathrm{P}(X_V \mid X_S = 1)$, by first applying the conditioning operation for $G$ to get $G_{|S}$ and then applying the classical ID-algorithm for conditional causal effect with latent variables on $G_{|S}$ (Shpitser and Pearl, 2006a).[24] This result seems to be new in the literature to our knowledge.[25] Similar generalizations can be made for other variants of the ID-algorithm, by first applying the conditioning operation for the graph and then applying the corresponding version of the ID-algorithm to the conditioned graph

[23] Note that in the usual c-ID-algorithm for conditional interventional distribution, the input is $\mathrm{P}(X_V)$ but not $\mathrm{P}(X_V \mid X_S = 1)$.

[24] If $T \cap \mathrm{Anc}_G(S) \neq \emptyset$, one can still apply the corresponding ID-algorithm to $G_{|S}$, but the algorithm would output an expression for $\mathrm{P}(X_A(x_T) \mid X_S = 1)$ instead of $\mathrm{P}(X_A \mid \mathrm{do}(X_T = x_T), X_S = 1)$. See Theorem 3.14 and Section 3.4.

[25] When we were writing this manuscript, we found that an s-ID-algorithm under latent variables was proposed in Abouei et al. (2024b). However, they only consider identification for the unconditional interventional distribution $\mathrm{P}(X_A \mid \mathrm{do}(X_T = x_T), X_S = 1)$, not for the conditional interventional distribution $\mathrm{P}(X_A \mid \mathrm{do}(X_T = x_T), X_B, X_S = 1)$.

(e.g., in the one-line formulation of the ID-algorithm Richardson et al. (2023, Theorem 48), replace $G$ with $G_{|S}$).

However, one should note that applying an ID-algorithm to the conditioned graph alone can hardly give a complete algorithm in general, due to the abstraction nature of the conditioning operation. For example, in the case of the s-ID-algorithm, we can use the conditioning operation to handle cases where $T \cap \mathrm{Anc}_G(S) = \emptyset$, but a complete algorithm should also be able to address cases where $T \cap \mathrm{Anc}_G(S) \neq \emptyset$ and $T \overset{d}{\underset{G_{\underline{T}}}{\perp}} A \mid S$ (see Abouei et al. (2024a, Theorem 1)).

## 4.3 Causal discovery under latent selection

Many causal discovery algorithms address unobserved common causes, but exclude selection bias. For simplicity, we consider consistent algorithms that output a single ADMG (instead of equivalence class). We can interpret the output of such algorithms as $G((M_{\backslash L})_{|\mathcal{S}})$ where $M$ is an acyclic SCM with latent nodes $L$, selection mechanism $X_S \in \mathcal{S}$, and $L \cap S = \emptyset$. This can give a certain causal interpretation to the output of these algorithms under selection bias even if selection bias is excluded in the original formulations. The high-level idea is: given one such algorithm $\mathcal{A}$, ideal infinite i.i.d. data $\mathcal{D}$ and a model class $\mathbb{M}$ of SCMs, the algorithm outputs a causal graph $\mathcal{A}(\mathcal{D})$ such that there exists $M$ in $\mathbb{M}$ such that $G(M) = \mathcal{A}(\mathcal{D})$. Now suppose that the data $\widetilde{\mathcal{D}}$ are generated by an s-SCM $M^S$ in some model class $\mathbb{M}^S$ such that the conditioning operation projects $\mathbb{M}^S$ into a subclass $\mathbb{M}_{|S}$ of $\mathbb{M}$. Then we can apply the *same algorithm* to get $\mathcal{A}(\widetilde{\mathcal{D}})$ and by the fact $\mathrm{P}_{M^S}(X_O) = \mathrm{P}_M(X_O \mid X_S \in \mathcal{S}) = \mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_O)$, we have $G(M_{|X_S \in \mathcal{S}}) = \mathcal{A}(\widetilde{\mathcal{D}})$. Theorems 3.14 and 3.17 and Proposition D.1 tell us that $\mathcal{A}(\widetilde{\mathcal{D}})$ can be seen as the closest approximation of $M^S$ in $\mathbb{M}$.

**Example 4.7** (Causal discovery)**.** For one instance, Wang and Drton (2023) explored recovering causal graphs uniquely from data generated by an acyclic linear non-Gaussian SCM with a bow-free graph (i.e., no simultaneous bidirected and directed edges between two variables) and rule out selection bias. Assume that the data are generated from an acyclic linear s-SCM $(M, X_S \in \mathcal{S})$ and the conditioned marginalized SCM of it has a bow-free graph. If the exogenous distribution of $(M_{\backslash L})_{|\mathcal{S}}$ is non-Gaussian and $(M_{\backslash L})_{|\mathcal{S}}$ satisfies the assumptions in Wang and Drton (2023, Section 3), then we can use the algorithm BANG in Wang and Drton (2023) to recover the graph of $(M_{\backslash L})_{|\mathcal{S}}$.

If we know from data or prior knowledge that a node $t$ is not an ancestor of $S$, then we can give a causal interpretation of $X_t$ in the discovered graph and apply causal identification results to identify $\mathrm{P}_M(X_O \mid \mathrm{do}(X_t = x_t), X_S \in \mathcal{S})$ with $O \subseteq V \setminus (L \cup S)$. For example, if the data are selected by $X_S = x_S$, we can sometimes read off whether $t \notin \mathrm{Anc}_{G(M)}(S)$ from a PAG (Partial Ancestral Graphs) or a MAG (Spirtes et al., 1995; Richardson and Spirtes, 2002).[26]

In addition to the causal discovery algorithms mentioned above, some causal model selection methods, such as the inflation technique (Wolfe et al., 2019), can also be generalized to deal with selection bias via the conditioning operation.

[26] Note that if $t \in \mathrm{Anc}_{G(M)}(S)$, we can still apply the identification result to the interventional distribution given $\mathrm{do}(X_t = x_t)$ in $M_{|X_S \in \mathcal{S}}$, but the causal identification results will output a formula for $\mathrm{P}_M(X_O(x_t) \mid X_S \in \mathcal{S})$ instead of $\mathrm{P}_M(X_O \mid \mathrm{do}(X_t = x_t), X_S \in \mathcal{S})$ (cf. Theorem 3.14, Section 3.4).

## 4.4 Instrumental variable and mediation analysis under latent selection

In some situations, we cannot achieve point identification results, but we can derive informative bounds for target causal effects. A well-known example is the instrumental inequality (Pearl, 2009; Balke and Pearl, 1994, 1997; Pearl, 1995b). More recent advances include, e.g., showing that the instrumental inequality is sharp for finite discrete variables under certain constraints on the cardinality of the variables (Badhane et al., 2025; Van Himbeeck et al., 2019), and extending the bounds to continuous outcomes (Zhang and Bareinboim, 2021). Not only can the original instrumental inequality for binary variables be extended to the case with certain selection bias immediately via the conditioning operation, but also the results we mentioned above.

**Example 4.8** (Instrumental variables)**.** The instrumental inequality was originally derived for the SCMs with the graph $G(M)$ shown in Figure 14. Similarly to Example 4.3, if we know that for an SCM $\widetilde{M}$ with latent variables $L$ and latent selection $S \in \mathcal{S}$, the causal graph $G\big(\big(\widetilde{M}_{\setminus L}\big)_{|\mathcal{S}}\big)$ takes the form shown in Figure 14, then we can conclude that the same form of inequality also holds for $\widetilde{M}$ under the subpopulation.

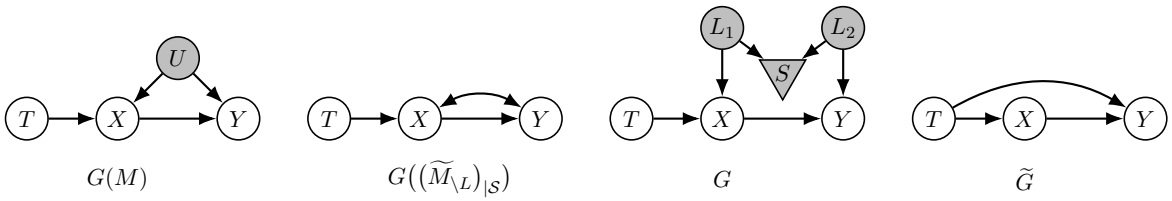



Figure 14: $G(M)$ and $G\big(\big(\widetilde{M}_{\setminus L}\big)_{|\mathcal{S}}\big)$ are graphs for the instrumental variables model. $G$ is the graph of a model with selection bias whose marginalized and conditioned graph is $G\big(\big(\widetilde{M}_{\setminus L}\big)_{|\mathcal{S}}\big)$ while $\widetilde{G}$ is its MAG representation.

If we further assume a continuous linear model $Y = \beta X + f(U)$ in $M$, then the parameter $\beta$ is identifiable when $\mathrm{Cov}(X, Y) \neq 0$ and is estimated as $\frac{\mathrm{Cov}_M(T,Y)}{\mathrm{Cov}_M(X,Y)}$, where selection bias is implicitly ruled out (Imbens et al., 2000). With the conditioning operation, we can see that the parameter remains identifiable from the selected conditional distribution $\mathrm{P}_{\widetilde{M}}(T, X, Y \mid S \in \mathcal{S})$ as $\frac{\mathrm{Cov}_{\widetilde{M}}(T,Y|S\in\mathcal{S})}{\mathrm{Cov}_{\widetilde{M}}(X,Y|S\in\mathcal{S})}$ even under certain forms of selection bias. Therefore, we have extended the identification result to include certain forms of selection bias.

**Remark 4.9.** Note that it is unclear how MAGs can handle this example. If graph $G$ in Figure 14 is interpreted as a MAG then conditioning on $S$ and marginalizing out $L_1$ and $L_2$ would yield the MAG $\widetilde{G}$ in Figure 14, where the assumptions of instrumental variables are violated, thus being too coarse to establish the instrumental inequality.

Mediation analysis is crucial in many fields such as epidemiology, natural science, and policy making, where understanding "path-specific" causal effects is often necessary (Pearl, 2001, 2014, 2009; Robins and Greenland, 1992). Traditional methods rely on linear regression, but linear SCMs have been proven problematic due to potential nonlinear interactions among variables, latent common causes, and selection bias in real-world problems (Shpitser, 2013).

**Example 4.10** (Mediation analysis)**.** With the help of potential outcomes and causal graphs of SCMs, Pearl (2014) and Shpitser (2013) study methods to perform mediation analysis when there are nonlinear functional dependencies and unobserved common causes. By extending the interpretation of bidirected edges to also represent *selection bias*, we can extend these results to account for selection bias immediately, similarly to the approach in previous examples.

For another example on how the conditioning operation is helpful, suppose that one is interested in the effect of, e.g., $A$ (obesity) on $Y$ (mortality) while conditioning a mediating variable on the path between them to a specific value (e.g., $S = 1$: having heart disease) (Smith, 2020). The graph $G$ is shown in Figure 15. Applying the graphical conditioning operation gives $G_{|S}$. This shows that we can obtain a causal identification result for $\mathrm{E}[Y(a) \mid S = 1] - \mathrm{E}[Y(a') \mid S = 1]$ via back-door adjustment on $L$.

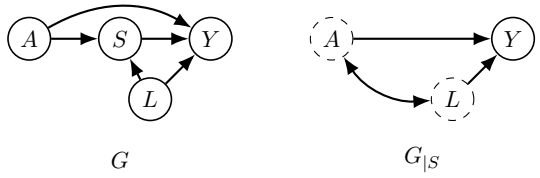



Figure 15: Graph $G$ for mediation analysis conditioning on one mediator and its conditioned graph $G_{|S}$.

**Remark 4.11.** The previous example shows that the conditioning operation is helpful in obtaining causal identification results, so it could also play a role in fairness analysis (Nabi and Shpitser, 2018; Chiappa, 2019; Kusner et al., 2017; Zhang and Bareinboim, 2018; Badhane et al., 2025).

## 4.5 Causal modeling under latent selection

The question of how to perform causal modeling under selection bias is one of the original main motivations for this work. In the following example, we show how the conditioning operation can help with causal modeling under (latent) selection bias. The high-level idea is from Example 4.1 that even if there are no causal effects and no common causes between two variables there could still be dependency between them caused by selection bias. To state the example, recall that one possible workflow of causal inference is:

(i) asking causal queries;

(ii) **building a causal model** from prior knowledge and data;

(iii) determining the target causal quantity and identifying the estimand in terms of available observational and interventional distributions;

(iv) using data to estimate the estimand.

As concise encodings of causal assumptions, causal graphs can be used to decide the estimand for addressing causal queries, and therefore incorrect graphs might generate wrong results.

**Example 4.12** (Causal modeling)**.** To understand the causal effect of treatment strategies from different countries on the fatality rate of COVID-19, Von Kügelgen et al. (2021) analyzed data from the initial virus outbreaks in 2020 in China and Italy, and assumed the causal graph $G$ shown in Figure 16. For COVID-19 infected people, age ($A$), country of residence ($C$) at the time of infection and fatality rate ($F$) are recorded.

The data suggest that $C$ and $A$ are dependent. In the traditional understanding of bidirected edges, assuming that $C$ and $A$ do not share a latent common cause, one has to draw a directed edge between $C$ and $A$ so that the hypothesized graph is compatible with the observation. However, drawing a directed edge from $C$ to $A$ is not a reasonable causal assumption. It assumes that if we conduct a randomized trial to assign people to different countries, then immediately ($A$ and $C$ are measured almost the same time) the resulting age distribution will differ depending on the assigned country. Similarly, $A \dashrightarrow C$ would also be an unreasonable assumption.

However, the conditioning operation tells us that bidirected edges do not have to represent latent common causes only, but can also represent latent selection bias. Therefore, we can draw a bidirected edge $C \dashleftarrow\dashrightarrow A$ as shown in $\widetilde{G}$ to explain the statistical association between $C$ and $A$, which could represent different latent selection mechanisms or latent common causes or combinations of the two between $C$ and $A$.[27] 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 the country and the age may influence the risk of getting infected, selection of the infection status ($S = 1$) can also lead to $C \dashleftarrow\dashrightarrow A$ (as in $G^2$). The combinations of both selection mechanisms (such 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 causal graphs in detail, including *all* relevant latent variables that model the selection mechanism. 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". It allows us to compute the total causal effect $\mathrm{TCE}(F; c' \to c) := \mathrm{E}[F \mid \mathrm{do}(C = c)] - \mathrm{E}[F \mid \mathrm{do}(C = c')]$ via the abstracted (conditioned) model $\widetilde{G}$ (e.g., by adjusting on age) without fully knowing all the latent details. Note that the results based on $G$ and $\widetilde{G}$ are clearly different. In fact, for an SCM with graph $G$, one has:

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

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

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

[27] Note that the difference of common cause and selection bias does not matter for the current task, which shows the power of model abstraction.

where in the second equality we use the back-door theorem allowed by the graph $\widetilde{G}$.

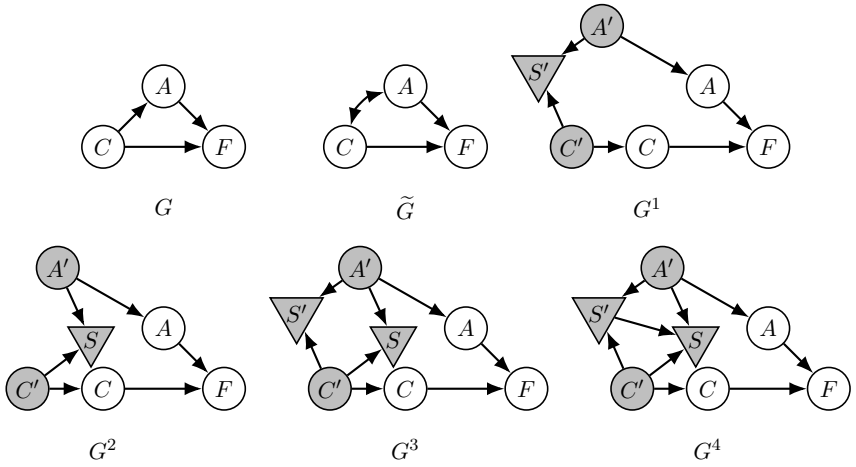



Figure 16: Causal graphs for COVID-19 data. Note that after applying the conditioning operation to selection variables and marginalizing out remaining latent variables, we reduce $G^i$ to $\widetilde{G}$ for $i = 1, 2, 3, 4$.

## 5 Discussion

Although SCMs have been widely used to study selection bias from a structural causal viewpoint, a formal theory was still absent. We gave a mathematical definition of s-SCMs (Definition 3.2), which formalizes the idea of SCMs with selection mechanisms, and a description of the data-generating processes that they are modeling. Motivated by the marginalization of causal models, which plays an important role in abstracting away unnecessary *unconditioned* latent details of causal models, we defined a conditioning operation (Definition 3.5) to transform an SCM with selection mechanisms into an SCM without selection mechanisms so that the new SCM preserves important information from the original SCM with selection mechanisms. The benefit of doing so is that, without the need to develop a separate theory for s-SCMs, we can *reduce* the problems involving s-SCMs to ordinary SCMs, so that all the well-developed tools from SCMs can be applied directly. We also explored the theoretical limit of such a transformation by showing what can be preserved (Section 3.2.2) and what is impossible to preserve (Appendix D.1) during such a model abstraction process.

Most importantly, we generalized the interpretation of bidirected edges in directed mixed graphs (interpreted as causal graphs of SCMs) so that they can represent not only latent common causes but also latent selection bias. This makes the rough idea of "using bidirected edges to represent selection bias" formal, such as Pearl's claim in his causality book (Pearl, 2009, p.163). Using the same symbol (bidirected edge) to represent potential latent common causes and latent selection bias is also consistent with some observation in epidemiology (Richardson and Robins, 2013a, Footnote 11). 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 under interventions, identification, and model selection.

One approach of causal modeling involves: (i) commencing with a complete graph, i.e., it has two directed edges in different directions and a bidirected edge between any two observed endogenous variables; (ii) iteratively deleting edges based on prior knowledge and available data. Our result contributes to this procedure by mathematically proving that, within the SCM framework, one should retain the bidirected edge between two variables when there is insufficient knowledge to rule out both unmeasured common cause and latent selection bias.

The current work focuses mainly on the theoretical aspects of the conditioning operation. Some of the applications are briefly examined. We envision exploring further and more detailed research of applications enabled by conditioning operation in future work. In particular, the conditioning operation might be helpful in giving a causal interpretation to the output of certain causal discovery algorithms under selection bias.

Markov categories have recently emerged as a categorical framework for probability and statistics (Fritz, 2020). In this "synthetic" approach, classical measure-theoretic foundation is replaced by a categorical one, and many familiar results can be proved algebraically within the framework (Fritz, 2020; Fritz et al., 2021; Chen et al., 2024a; Fritz et al., 2025). Crucially, causal modeling can also be formulated at this abstract level (Fritz and Klingler, 2023; Lorenz and Tull, 2023). In particular, it is possible to extend the theory of conditioning SCMs to the categorical setting, where recent work on partializations of Markov categories might be relevant (Mohammed, 2025).

## Acknowledgments and Disclosure of Funding

The authors acknowledge Booking.com for support. The authors thank Philip Boeken, Stephan Bongers, Robin Evans, Tobias Fritz, Areeb Shah Mohammed, and Thomas Richardson for discussions. The authors are grateful to Luigi Gresele for discussions regarding Example 4.12 and for affirming our approach to causal modeling of the Covid example.

# 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 A.1** (Twin SCM)**.** *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)***

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

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

**Definition A.2** (Parent)**.** *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 $\widetilde{f}_v : \mathcal{X}_{V\setminus k} \times \mathcal{X}_{W\setminus k} \to \mathcal{X}_v$ such that for $\mathrm{P}(X_W)$-almost every $x_W \in \mathcal{X}_W$, for all $x_V \in \mathcal{X}_V$,*

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

**Definition A.3** (Graph and augmented graph)**.** *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 $v \in V$.*

Note that $G(M) = (G^a(M))_{\setminus W}$, where the graphical marginalization (also known as "latent projection") is defined in Bongers et al. (2021, Definition 5.7).

**Example A.4.** 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 17.

**Definition A.5** ((Counterfactual/interventional/observational) equivalence)**.** *A simple SCM $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ is* ***counterfactually equivalent*** *to a simple SCM $\widetilde{M} = (\widetilde{V}, \widetilde{W}, \widetilde{\mathcal{X}}, \widetilde{\mathrm{P}}, \widetilde{f})$ w.r.t. $O \subseteq V \cap \widetilde{V}$ if for any $T_1 \subseteq O$ and $x_{T_1} \in \mathcal{X}_{T_1}$, and any $T_2 \subseteq O'$ and $x_{T_2} \in \mathcal{X}_{T_2}$,*

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

*We say $M$ is* ***interventionally equivalent*** *to $\widetilde{M}$ w.r.t. $O$ if*

$$\mathrm{P}_M(X_{O\setminus T_1} \mid \mathrm{do}(X_{T_1} = x_{T_1})) = \mathrm{P}_{\widetilde{M}}(X_{O\setminus T_1} \mid \mathrm{do}(X_{T_1} = x_{T_1})).$$

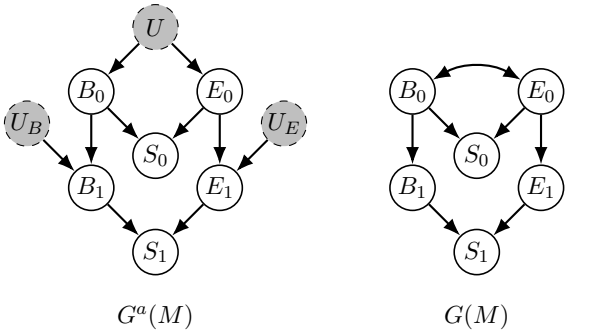



Figure 17: The (augmented) causal graphs of the SCM $M$ in Example A.4.

*We say $M$ is **observationally equivalent** to $\widetilde{M}$ w.r.t. $O$ if*

$$\mathrm{P}_M(X_O) = \mathrm{P}_{\widetilde{M}}(X_O).$$

*We say that $M$ and $\widetilde{M}$ are observationally/interventionally/counterfatually equivalent if $V = \widetilde{V}$ and $M$ is observationally/interventionally/counterfatually equivalent to $\widetilde{M}$ w.r.t. $V$.*

**Remark A.6.** We have

$$\begin{aligned} \text{Equivalence of SCMs} &\implies \text{Counterfactual equivalence} \\ &\implies \text{Interventional equivalence} \\ &\implies \text{Observational equivalence,} \end{aligned}$$

but not conversely. See Bongers et al. (2021, Proposition 4.6).

**Definition A.7** (Directed global Markov property)**.** *Let $G$ be a DMG with nodes $V$ and $\mathrm{P}(X_V)$ a probability distribution on $\mathcal{X}_V = \prod_{v\in V} \mathcal{X}_v$ for standard measurable spaces $X_v$. We say that the probability distribution $\mathrm{P}(X_V)$ satisfies the **directed global Markov property relative to** $G$ if for subsets $A, B, C \subseteq V$ the set $A$ being d-separated from $B$ given $C$ implies that the random variable $X_A$ is conditionally independent of $X_B$ given $X_C$.*

**Theorem A.8** (Directed Markov property for SCMs; Forré and Mooij (2017))**.** *Let $M$ be a uniquely solvable SCM that satisfies at least one of the following three conditions:*

*(1) $M$ is acyclic;*

*(2) all endogenous state spaces $\mathcal{X}_v$ are discrete and $M$ is ancestrally uniquely solvable (Bongers et al., 2021, Definition 3.9);*

*(3) $M$ is linear (Bongers et al., 2021, Definition C.1) and each of its causal mechanisms $\{f_v\}_{v\in V}$ has a nontrivial dependence on at least one exogenous variable, and $\mathrm{P}(X_W)$ has a density w.r.t. the Lebesgue measure on $\mathbb{R}^W$.*

*Then its observational distribution $\mathrm{P}_M(X_V)$ exists, is unique, and satisfies the directed global Markov property relative to $G(M)$.*

We first recall the definition of $\sigma$-separation. In the following, we write

$$\mathrm{Sc}_G(C) \coloneqq \{\widetilde{v} \in V : \exists \widetilde{v} \dashrightarrow \cdots \dashrightarrow v \text{ and } \widetilde{v} \dashleftarrow \cdots \dashleftarrow v \text{ for some } v \in C\}$$

for the **strongly connected component** of $C \subseteq V$.

**Definition A.9** ($\sigma$-sepation for DMGs, (Forré and Mooij, 2017; Bongers et al., 2021))**.** *Let $G$ be a DMG with nodes $V$ and $C \subseteq V$ a subset of nodes and $\pi$ a walk in $G$:*

$$\pi = (v_0 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast v_n)\,.$$

*(1) We say that the walk $\pi$ is $C$-$\sigma$-**blocked** or $\sigma$-**blocked by** $C$ if:*

*(i) $v_0 \in C$ or $v_n \in C$ or;*

*(ii) there are two adjacent edges in $\pi$ of one of the following forms:*

$$\begin{aligned}
&\textit{left chain:} && v_{k-1} \dashleftarrow v_k \ast\!\!-\!\!\ast v_{k+1} && \textit{with} && v_k \in C \wedge v_k \notin \mathrm{Sc}_G(v_{k-1})\,,\\
&\textit{right chain:} && v_{k-1} \ast\!\!-\!\!\ast v_k \dashrightarrow v_{k+1} && \textit{with} && v_k \in C \wedge v_k \notin \mathrm{Sc}_G(v_{k+1})\,,\\
&\textit{fork:} && v_{k-1} \dashleftarrow v_k \dashrightarrow v_{k+1} && \textit{with} && v_k \in C \wedge v_k \notin \mathrm{Sc}_G(v_{k-1}) \cap \mathrm{Sc}_G(v_{k+1})\,,\\
&\textit{collider:} && v_{k-1} \ast\!\!-\!\!\ast v_k \dashleftarrow \ast\!\!-\!\!\ast v_{k+1} && \textit{with} && v_k \notin \mathrm{Anc}_G(C).
\end{aligned}$$

*We say that the walk $\pi$ is $C$-$\sigma$-open if it is not $C$-$\sigma$-blocked.*

*(2) Let $A, B, C \subseteq V$ (not necessarily disjoint) be subsets of nodes. We then say that: $A$ **is $\sigma$-separated from $B$ given $C$ in $G$**, in symbols:*

$$A \overset{\sigma}{\underset{G}{\perp}} B \mid C,$$

*if every walk/path from a node in $A$ to a node in $B$ is $\sigma$-blocked by $C$. (In the definition, taking either walk or path gives an equivalent definition.)*

**Definition A.10** (Generalized directed global Markov property; Forré and Mooij (2017))**.** *Let $G = (V, E, H)$ be a DMG and $\mathrm{P}(X_V)$ a probability distribution on $\mathcal{X}_V = \prod_{v \in V} \mathcal{X}_v$ for standard measurable spaces $X_v$. We say that the probability distribution $\mathrm{P}(X_V)$ satisfies the **generalized directed global Markov property relative to** $G$ if for subsets $A, B, C \subseteq V$ the set $A$ being $\sigma$-separated (Definition A.9) from $B$ given $C$ implies that the random variable $X_A$ is conditionally independent of $X_B$ given $X_C$.*

**Theorem A.11** (Generalized directed Markov property for SCMs; Forré and Mooij (2017); Bongers et al. (2021))**.** *Let $M$ be a simple SCM. Then its observational distribution $\mathrm{P}_M(X_V)$ exists, is unique, and satisfies the generalized directed global Markov property relative to $G(M)$.*

# B Some examples and remarks

**Remark B.1** (Proof of claim in Remark 2.13)**.** If we assume that there is an underlying acyclic SCM $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ inducing the potential outcomes $X_A$ and $X_B(x_A)$, then equation (1) is equivalent to

$$\forall x_A \in \{0,1\} : g_A(X_W) \perp\!\!\!\perp g_B^{V \setminus A}(x_A, X_W), \tag{5}$$

where $g$ and $g^{V\setminus A}$ are the (essentially unique) solution functions of $M$ w.r.t. $V$ and $V \setminus A$, respectively. Then we can show that (under a positivity assumption) equation (2) holds.

Indeed, for every $x_A \in \{0,1\}$ with $\mathrm{P}_M(X_A = x_A) > 0$,

$$\begin{aligned}
\mathrm{P}_M(X_B \mid X_A = x_A) &= \mathrm{P}_M(g_B(X_W) \mid g_A(X_W) = x_A) \\
&= \mathrm{P}_M(g_B^{V\setminus A}(g_A(X_W), X_W) \mid g_A(X_W) = x_A) && \text{(Lemma C.3)} \\
&= \mathrm{P}_M(g_B^{V\setminus A}(x_A, X_W) \mid g_A(X_W) = x_A) \\
&= \mathrm{P}_M(g_B^{V\setminus A}(x_A, X_W)) && \text{(equation (5))} \\
&= \mathrm{P}_M(X_B(x_A)) \\
&= \mathrm{P}_M(X_B \mid \mathrm{do}(X_A = x_A)).
\end{aligned}$$

**Remark B.2** (Assumption 3.1 is mild)**.** (1) Note that the class of simple SCMs is a *more general model class* than acyclic SCMs. Besides, one can easily generalize all the results in this work to an even more general class of SCMs than simple SCMs by carefully postulating corresponding unique solvability assumptions for the SCM w.r.t. certain subsets of $V$. However, for non-simple SCMs, there are many counter-intuitive phenomena. For example, non-simple SCMs can induce none or multiple (observational and interventional) distributions, they lack a Markov property, and interventions may affect non-descendants of the intervened target (see e.g., Bongers et al. (2021)). This suggests that non-simple SCMs are not intuitive causal models. Therefore, we focus on simple SCMs in the current work.

(2) In real-world applications, we mostly observe data from events with positive probabilities. Also, mathematically, although measure theory provides a way to define conditional probabilities given a null event, it is still ambiguous in general when the Borel-Kolmogorov paradox arises (**?**Jaynes, 2003). So, it is reasonable not to model selection events with zero probabilities.

**Remark B.3** (Remark on Proposition 3.29)**.** (1) Recall that $G(M_{\setminus L})$ can be a strict subgraph of $G(M)_{\setminus L}$. Also note that the "merging step" can be more coarse in $G(M)$ than in $M$. Therefore, $G(M_{|X_S \in \mathcal{S}})$ can be a strict subgraph of $G(M)_{|S}$ due to the merging step and the marginalization. This means that $G(M)_{|S}$ is generally a (strictly) more conservative representation of the underlying conditioned SCM with less causal information due to the nature of abstraction (recall that a sparser causal graph encodes stronger assumptions).

(2) Any reasonable attempt to a purely graphical conditioning operation should satisfy this property. Otherwise, one would conclude from the graph $G_{|S}$ some results that do not hold for some conditioned SCMs $M_{|X_S \in \mathcal{S}}$ where $M$ is compatible with the graph $G$. Also note that one is not able to further "minimize" the conditioned graph by eliminating some (bi)directed edges, since there always exists an SCM $M$ and a selection mechanism $X_S \in \mathcal{S}$ such that $G(M_{|X_S \in \mathcal{S}}) = G(M)_{|S}$ (consider a linear SCM with positive coefficients in which every endogenous variable has at least one exogenous parent such that $\mathrm{P}(X_w) = \mathcal{N}(0,1)$ for all $w \in W$ and with selection mechanism $X_S \in [0, \infty)$).

(3) Recall that graphical marginalization preserves ancestral relationships, i.e., $\mathrm{Anc}_{G_{\setminus L}}(B) = \mathrm{Anc}_G(B) \setminus L$. This property also holds for graphical conditioning, i.e., $a \in \mathrm{Anc}_G(b)$ iff

$a \in \mathrm{Anc}_{G_{|S}}(b)$ for any $S \subseteq V \setminus \{a, b\}$. However, this is not the case for SCM marginalization and conditioning in general. At the level of SCMs, the best we can conclude is that if $a \in \mathrm{Anc}_{G(M_{\setminus L})}(b)$ or $a \in \mathrm{Anc}_{G(M_{|X_S \in \mathcal{S}})}(b)$, then $a \in \mathrm{Anc}_{G(M)}(b)$ but not conversely.

(4) The conditioning operation for an SCM does not introduce new directed causal paths to the graph of the original SCM. This aligns with the principle that an "individual causal effect" present in a subset of the population must also be present in the entire population, though the reverse is not necessarily true.[28]

**Remark B.4** (Remark on modeling interpretation)**.** (1) The "node-splitting" trick has been applied in various forms and situations (Pearl, 2009; Richardson and Robins, 2013b). Usually, a copy $A'$ is added as a child of $A$ to the original causal model and interventions are performed on $A'$ instead of on $A$. Similarly, one can also use the node-splitting trick when using the conditioning operation where a copy $A'$ is added as a parent of $A$ to the original causal model and selections are performed on $A'$ instead of on $A$.

(2) There are some relations between conditioned SCMs and counterfactual reasoning. Therefore, the conditioning operation can provide an easy way to identify *unnested* counterfactual quantities in some cases.[29] For example, suppose that we want to identify the counterfactual quantity $\mathrm{P}_M(Y(t) \mid S = s)$ given the graph $G(M)$ in Figure 18. Then we can apply the graphical conditioning operation to get $G(M)_{|S}$ from $G(M)$. We know that $G(M_{|S=s})$ must be a subgraph of $G(M)_{|S}$ by Proposition 3.29. For finite discrete variables under positivity assumptions, the back-door criterion applied to $M_{|S=s}$ gives

$$\mathrm{P}_{M_{|S=s}}(Y(t)) = \sum_z \mathrm{P}_{M_{|S=s}}(Y \mid T = t, Z = z)\mathrm{P}_{M_{|S=s}}(Z = z).$$

Hence, we can conclude that

$$\mathrm{P}_M(Y(t) \mid S = s) = \sum_z \mathrm{P}_M(Y \mid T = t, Z = z, S = s)\mathrm{P}_M(Z = z \mid S = s).$$

This is related to the problems of "Type-I selection bias" and "internal validity" according to the jargon of Lu et al. (2022) and Smith (2020), respectively.

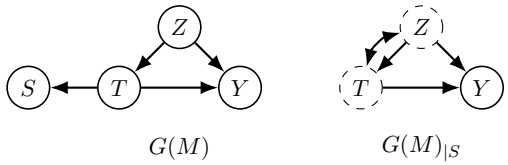



Figure 18: Causal graph $G(M)$ and its conditioned graph $G(M)_{|S}$ on $S$.

[28] Note that in contrast to "individual level" causal effects, a "population causal effect" in a subpopulation may not be present in the whole population due to the fact that "population causal effects" in different subpopulations may cancel out with each other.

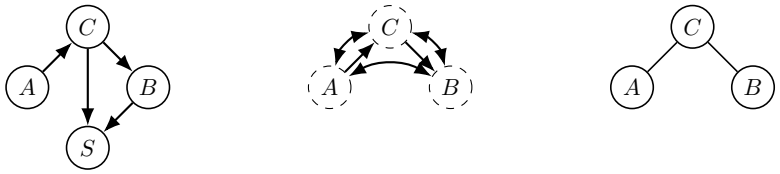
[29] Notions defined via nested counterfactual quantity such as various notions of fairness (Kusner et al., 2017; Zhang and Bareinboim, 2018) can always be rewritten as an unnested one using the Counterfactual Unnesting Theorem (Correa et al., 2021, Theorem 1).

**Example B.5** (The assumptions in Theorem 3.25 cannot be omitted)**.** (1) The assumption that $\mathrm{Ch}_G(S) = \emptyset$ cannot be omitted. Consider the following case shown in Figure 19. We have $A \overset{\sigma}{\underset{G}{\perp}} B \mid S$ but do not have $A \overset{\sigma}{\underset{G_{|S}}{\perp}} B$.


$G$ $\quad$ $G_{|S}$

Figure 19: Causal graphs in the first item of Example B.5.

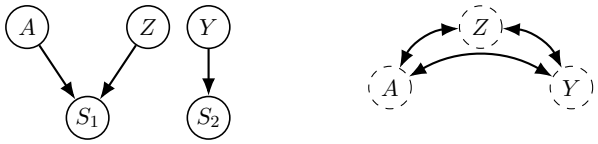
(2) The assumption that $C \cap \mathrm{Anc}_G(S) = \emptyset$ cannot be omitted. Consider the following case shown in Figure 20. We have $A \overset{\sigma}{\underset{G}{\perp}} B \mid C \cup S$ but we do not have $A \overset{\sigma}{\underset{G_{|S}}{\perp}} B \mid C$. However, note that we have $A \overset{d}{\underset{\mathrm{MAG}(G)_{|S}}{\perp}} B \mid C$ where $\mathrm{MAG}(G)_{|S}$ denotes the conditioned MAG of $G$ given $S$ (see Richardson and Spirtes (2002)).




$G$ $\quad$ $G_{|S}$ $\quad$ $\mathrm{MAG}(G)_{|S}$

Figure 20: Causal graphs and MAGs in the second item of Example B.5.

(3) The assumption that $S$ is a singleton set cannot be omitted. Consider the following case shown by Figure 21. For $S = \{S_1, S_2\}$, we have $A \overset{\sigma}{\underset{G}{\perp}} Y \mid Z \cup S$, but we do not have $A \overset{\sigma}{\underset{G_{|S}}{\perp}} Y \mid Z$.



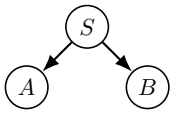

$G$ $\quad$ $G_{|S}$

Figure 21: When $S = \{S_1, S_2\}$ is not a singleton set, the $\sigma$-separation $A \overset{\sigma}{\underset{G}{\perp}} Y \mid Z \cup S$ does not imply $A \overset{\sigma}{\underset{G_{|S}}{\perp}} Y \mid Z$.

**Example B.6** (Markov property does not imply conditional independence given an event)**.** Consider a discrete acyclic causal model $M$ given by

$$\begin{aligned}
&\mathrm{P}_M(X_A \mid X_S = 0) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1, \quad \mathrm{P}_M(X_B \mid X_S = 0) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1,\\
&\mathrm{P}_M(X_A \mid X_S = 1) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1, \quad \mathrm{P}_M(X_B \mid X_S = 1) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1,\\
&\mathrm{P}_M(X_A \mid X_S = 2) = \frac{1}{3}\delta_0 + \frac{2}{3}\delta_1, \quad \mathrm{P}_M(X_B \mid X_S = 2) = \frac{1}{3}\delta_0 + \frac{2}{3}\delta_1,\\
&\mathrm{P}_M(X_S) = \frac{1}{3}\delta_0 + \frac{1}{3}\delta_1 + \frac{1}{3}\delta_2,
\end{aligned}$$

and $\mathrm{P}_M(X_A, X_B, X_S) = \mathrm{P}_M(X_A \mid X_S) \otimes \mathrm{P}_M(X_B \mid X_S) \otimes \mathrm{P}_M(X_S)$ whose graph is drawn in Figure 22.




$G(M)$

Figure 22: Causal graph of $M$ where $X_A \not\perp\!\!\!\perp_{\mathrm{P}_M(X_V)} X_B \mid X_S \in \{1,2\}$ even if $A \overset{d}{\perp}_{G(M)} B \mid S$.

We can compute that

$$\begin{aligned}
\mathrm{P}_M(X_A \mid X_S \in \{1,2\}) &= \frac{\mathrm{P}_M(X_A, X_S \in \{1,2\})}{\mathrm{P}_M(X_S \in \{1,2\})} = \frac{5}{12}\delta_0 + \frac{7}{12}\delta_1,\\
\mathrm{P}_M(X_B \mid X_S \in \{1,2\}) &= \frac{\mathrm{P}_M(X_B, X_S \in \{1,2\})}{\mathrm{P}_M(X_S \in \{1,2\})} = \frac{5}{12}\delta_0 + \frac{7}{12}\delta_1,\\
\mathrm{P}_M(X_A, X_B \mid X_S \in \{1,2\}) &= \frac{\mathrm{P}_M(X_A, X_B, X_S \in \{1,2\})}{\mathrm{P}_M(X_S \in \{1,2\})} = \frac{13}{72}\delta_{00} + \frac{17}{72}\delta_{01} + \frac{17}{72}\delta_{10} + \frac{25}{72}\delta_{11}.
\end{aligned}$$

Then it is easy to see that

$$\begin{aligned}
&\mathrm{P}_M(X_A \mid X_S \in \{1,2\}) \otimes \mathrm{P}_M(X_B \mid X_S \in \{1,2\})\\
&= \frac{25}{144}\delta_{00} + \frac{35}{144}\delta_{01} + \frac{35}{144}\delta_{10} + \frac{49}{144}\delta_{11}\\
&\neq \mathrm{P}_M(X_A, X_B \mid X_S \in \{1,2\}).
\end{aligned}$$

Note that $A \overset{d}{\perp}_{G(M)} B \mid S$. From the Markov property or direct calculation, one can see that $X_A \perp\!\!\!\perp_{\mathrm{P}_M(X_V)} X_B \mid X_S$. However, the above calculation implies that $X_A \not\perp\!\!\!\perp_{\mathrm{P}_M(X_V)} X_B \mid X_S \in \{1,2\}$.

**Remark B.7** (Remark on Example B.6)**.** Note that one *cannot* say that this falsifies the Markov property. In fact, the Markov property is about conditioning on a variable and the above example is about conditioning on an event, which are fundamentally different.

# C Proofs

**Lemma 3.4** (Finest partition)**.** *Let $\mathfrak{P}_{\mathcal{S}}$ denote the set of partitions $\mathcal{I} = \{I_1, \dots, I_p\}$ of $W$ such that $\{X_{I_i}\}_{i=1}^p$ are mutually independent under $\widetilde{\mathrm{P}}(X_W) = \mathrm{P}(X_W \mid X_S \in \mathcal{S})$. Then there exists $\mathcal{H} \in \mathfrak{P}_{\mathcal{S}}$ such that $\mathcal{H}$ is a finer partition than any other partition $\mathcal{I} \in \mathfrak{P}_{\mathcal{S}}$.*

*Proof.* We first show that $(\mathfrak{P}_{\mathcal{S}}, \vee)$ is a finite join semi-lattice where $\mathcal{I} \vee \mathcal{J} \coloneqq \{I \cap J : I \in \mathcal{I} \text{ and } J \in \mathcal{J}\} \setminus \{\emptyset\}$. To achieve that, it suffices to show that $\mathfrak{P}_{\mathcal{S}}$ is closed under the join operation. If $\{X_{I_i}\}_{i=1}^p$ are mutually independent and $\{X_{J_j}\}_{j=1}^q$ are mutually independent under some probability distribution $\widetilde{\mathrm{P}}$ then we have that $\{X_{K_k}\}_{k=1}^m$ are mutually independent under $\widetilde{\mathrm{P}}$ where $\{K_k\}_{k=1}^m = \mathcal{I} \vee \mathcal{J}$. That is, $\widetilde{\mathrm{P}}(X_W) = \bigotimes_{i=1}^p \widetilde{\mathrm{P}}(X_{I_i}) = \bigotimes_{i=1}^p \bigotimes_{j=1}^q \widetilde{\mathrm{P}}(X_{I_i \cap J_j}) = \bigotimes_{k=1}^m \widetilde{\mathrm{P}}(X_{K_k})$. Since $(\mathfrak{P}_{\mathcal{S}}, \vee)$ is finite, there must exist a largest element, which we denote by $\mathcal{H} = \{H_i\}_{i=1}^n$. This means that every partition $\mathcal{J}$ in $\mathfrak{P}_{\mathcal{S}}$ must be coarser than $\mathcal{H}$ according to the order induced by the join $\vee$, so the partition $\mathcal{H}$ is the finest partition in $\mathfrak{P}_{\mathcal{S}}$. □

**Proposition 3.10** (Simple, acyclic, linear SCMs and conditioning)**.** *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.* Since exogenous random variables do not have parents, merging exogenous random variables will not introduce cycles. Merging exogenous random variables will also preserve the simplicity and linearity of SCMs. Indeed, if $g^A : \mathcal{X}_{V \setminus A} \times \mathcal{X}_W \to \mathcal{X}_A$ is the essentially unique solution function of $M$ w.r.t. $A$ for some $A \subseteq V \setminus S$, then the function $\widetilde{g}^A : \mathcal{X}_{V \setminus A} \times \mathcal{X}_{\widehat{W}} \to \mathcal{X}_A$ defined by $\widetilde{g}^A(x_{V \setminus A}, x_{\widehat{W}}) = g^A(x_{V \setminus A}, x_W)$ with $x_W = (x_{\widehat{w}})_{\widehat{w} \in \widehat{W}} = x_{\widehat{W}} \in \mathcal{X}_{\widehat{W}}$ is the essentially unique solution function of $\widetilde{M}$ w.r.t. $A$, where $\widetilde{M}$ is the same as $M$ but with $W$ replaced by $\widehat{W}$ and $\mathcal{X}$ by $\mathcal{X}_V \times \mathcal{X}_{\widehat{W}}$. So merging exogenous variables preserves simplicity. For linearity, let $\mathcal{X}_u$ and $\mathcal{X}_v$ denote some linear vector spaces (they do not have to be the real line $\mathbb{R}$), and $\mathcal{L}(\mathcal{X}_u, \mathcal{X}_v)$ denote the set of linear mappings from $\mathcal{X}_u$ to $\mathcal{X}_v$. Let $f_v(x_V, x_W) = \sum_{u \in V} (T_{vu}(x_W)) x_u + \Gamma_v(x_W)$ where $T_{vu} : \mathcal{X}_W \to \mathcal{L}(\mathcal{X}_u, \mathcal{X}_v)$ and $\Gamma_v : \mathcal{X}_W \to \mathcal{X}_v$ are (nonlinear) mappings such that $f_v$ is measurable. Then the mapping $\widetilde{f}_v(x_V, x_{\widehat{W}}) = \sum_{u \in V} (\widetilde{T}_{vu}(x_{\widehat{W}})) x_u + \widetilde{\Gamma}_v(x_{\widehat{W}})$ is still linear and measurable where

$$\widetilde{T}_{vu}(x_{\widehat{W}}) \coloneqq T_{vu}(x_W) \text{ and } \widetilde{\Gamma}_v(x_{\widehat{W}}) \coloneqq \Gamma_v(x_W) \text{ with } x_{\widehat{W}} = x_W.$$

Updating the probability distributions of the exogenous random variables to the posterior preserves simplicity, acyclicity, and linearity of SCMs. In particular, this preserves simplicity because if $\mathrm{P}_M(X_S \in \mathcal{S}) > 0$, then $\mathrm{P}_M(X_W \mid X_S \in \mathcal{S}) \ll \mathrm{P}_M(X_W)$. By slightly generalizing Bongers et al. (2021, Propositions 8.2, 5.11 and C.5),[30] we have that marginalization preserves simplicity, acyclicity, and linearity of SCMs. Hence, we obtain that the conditioning operation preserves simplicity, acyclicity, and linearity of SCMs. □

**Lemma 3.13** (Conditioning and intervention)**.** *Assume Assumption 3.1. Let $T \subseteq V \setminus \mathrm{Anc}_{G(M)}(S)$ and $x_T \in \mathcal{X}_T$. Then we have*

$$\left(M_{\mathrm{do}(X_T = x_T)}\right)_{|X_S \in \mathcal{S}} \equiv \left(M_{|X_S \in \mathcal{S}}\right)_{\mathrm{do}(X_T = x_T)}.$$

[30] Our definition of linear SCMs is more general than the one in Bongers et al. (2021).

*Proof.* We check the definition one by one. Write $\left(M_{|\mathcal{S}}\right)_{\mathrm{do}(X_T=x_T)} \coloneqq (\widehat{V}, \widehat{W}, \widehat{\mathcal{X}}, \widehat{\mathrm{P}}, \widehat{f})$ and $\left(M_{\mathrm{do}(X_T=x_T)}\right)_{|\mathcal{S}} \coloneqq (\overline{V}, \overline{W}, \overline{\mathcal{X}}, \overline{\mathrm{P}}, \overline{f})$. Set $O \coloneqq V \setminus S$.

First, it is easy to see that $\widehat{V} = V \setminus S = \overline{V}$. Because $T \cap \mathrm{Anc}_{G(M)}(S) = \emptyset$ and $M$ is simple, $g_S^{-1}(\mathcal{S}) = \widetilde{g}_S^{-1}(\mathcal{S})$ up to a $\mathrm{P}(X_W)$-null set where $g$ and $\widetilde{g}$ are solution functions of $M$ and $M_{\mathrm{do}(X_T=x_T)}$ respectively. Therefore, we can conclude that $\widehat{W} = \overline{W}$. Then we have $\widehat{\mathcal{X}} = \mathcal{X}_O \times \mathcal{X}_{\widehat{W}} = \mathcal{X}_O \times \mathcal{X}_{\overline{W}} = \overline{\mathcal{X}}$. Since $M$ and $M_{\mathrm{do}(X_T=x_T)}$ have the same exogenous distribution $\mathrm{P}$ and $g_S^{-1}(\mathcal{S}) = \widetilde{g}_S^{-1}(\mathcal{S})$ up to a $\mathrm{P}(X_W)$-null set, we have $\mathrm{P}_M(X_{\widehat{w}_i}, X_S \in \mathcal{S}) = \mathrm{P}_{M_{\mathrm{do}(X_T=x_T)}}(X_{\overline{W}_i}, X_S \in \mathcal{S})$. Hence, we can conclude that

$$\mathrm{P}_M(X_{\widehat{w}_i} \mid X_S \in \mathcal{S}) = \mathrm{P}_{M_{\mathrm{do}(X_T=x_T)}}(X_{\overline{W}_i} \mid X_S \in \mathcal{S}).$$

Therefore, we have $\widehat{\mathrm{P}} = \bigotimes_{i=1}^n \widehat{\mathrm{P}}(X_{\widehat{w}_i}) = \bigotimes_{i=1}^n \overline{\mathrm{P}}(X_{\overline{W}_i}) = \overline{\mathrm{P}}$. For the causal mechanisms, we have $\widehat{f}\left(x_{\widehat{V}}, x_{\widehat{W}}\right) = \left(f_{O\setminus T}\left(x_O, g^S\left(x_O, x_{\widehat{W}}\right), x_{\widehat{W}}\right), x_T\right)$ with $g^S$ the (essentially unique) solution function of $M$ w.r.t. $S$. Let $\widetilde{f}$ be the causal mechanism of $M_{\mathrm{do}(X_T=x_T)}$ and let $\widetilde{g}^S$ be the (essentially unique) solution function of $M_{\mathrm{do}(X_T=x_T)}$ w.r.t. $S$. Then we have

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

Since $T \cap S = \emptyset$, we have $f_S(x_V, x_W) = \widetilde{f}_S(x_V, x_W)$ for all $x_W \in \mathcal{X}_W$ and all $x_V \in \mathcal{X}_V$. Recall that $g^S$ and $\widetilde{g}^S$ are the (essentially unique) solution functions of $M$ and $M_{\mathrm{do}(X_T=x_T)}$ w.r.t. $S$ respectively, i.e., for $\mathrm{P}(X_W)$-a.a. $x_W \in \mathcal{X}_W$ and all $x_V \in \mathcal{X}_V$

$$x_S = g^S(x_O, x_W) \iff x_S = f_S(x_V, x_W) \text{ and } x_S = \widetilde{g}^S(x_O, x_W) \iff x_S = \widetilde{f}_S(x_V, x_W).$$

Therefore, for $\mathrm{P}(X_W \mid X_S \in \mathcal{S})$-a.a. $x_{\widehat{W}} = x_{\overline{W}} \in \mathcal{X}_{\widehat{W}} = \mathcal{X}_{\overline{W}}$ and all $x_O \in \mathcal{X}_O$

$$g^S(x_O, x_{\widehat{W}}) = \widetilde{g}^S(x_O, x_{\overline{W}}).$$

Hence, for $\mathrm{P}(X_W \mid X_S \in \mathcal{S})$-a.a. $x_{\widehat{W}} = x_{\overline{W}} \in \mathcal{X}_{\widehat{W}} = \mathcal{X}_{\overline{W}}$ and all $x_O \in \mathcal{X}_O$

$$\widehat{f}(x_O, x_{\widehat{W}}) = \overline{f}(x_O, x_{\overline{W}}).$$

With the definition of the conditioning operation we conclude

$$\left(M_{\mathrm{do}(X_T=x_T)}\right)_{|X_S\in\mathcal{S}} \equiv \left(M_{|X_S\in\mathcal{S}}\right)_{\mathrm{do}(X_T=x_T)}.$$

□

**Theorem 3.14** (Main result I: Causal semantics of conditioned SCMs)**.** *Assume Assumption 3.1 and write $O \coloneqq V \setminus S$. Let $T_i \subseteq O$ and $x_{T_i} \in \mathcal{X}_{T_i}$ for $i = 1, \ldots, n$. Then we have*

$$\mathrm{P}_{M_{|X_S\in\mathcal{S}}}\left(\{X_{O\setminus T_i}(x_{T_i})\}_{1\le i\le n}\right) = \mathrm{P}_M\left(\{X_{O\setminus T_i}(x_{T_i})\}_{1\le i\le n} \mid X_S \in \mathcal{S}\right).$$

*Proof.* Let $T \subseteq O$. Let $\widetilde{g}^{O\setminus T} : \mathcal{X}_T \times \mathcal{X}_W \to \mathcal{X}_{O\setminus T}$ be the (essentially unique) solution function of $M_{\setminus S}$ w.r.t. $O \setminus T$ and $\widehat{g}^{O\setminus T} : \mathcal{X}_T \times \mathcal{X}_{\widehat{W}} \to \mathcal{X}_{O\setminus T}$ be the (essentially unique) solution function of $M_{|\mathcal{S}}$ w.r.t. $O \setminus T$. For $\mathrm{P}_M(X_W \mid X_S \in \mathcal{S})$-a.a $x_{\widehat{W}} = x_W \in \mathcal{X}_W$ and all $x_V \in \mathcal{X}_V$, we have

$$\begin{aligned} x_{O\setminus T} = \widetilde{g}^{O\setminus T}(x_T, x_W) &\iff x_{O\setminus T} = f_{O\setminus T}\left(x_O, g^S(x_O, x_W), x_W\right) \\ &\iff x_{O\setminus T} = \widehat{f}_{O\setminus T}(x_O, x_W) \\ &\iff x_{O\setminus T} = \widehat{g}^{O\setminus T}\left(x_T, x_{\widehat{W}}\right) \end{aligned}$$

This implies that $\widetilde{g}^{O\setminus T}(x_W, x_T) = \widehat{g}^{O\setminus T}(x_{\widehat{W}}, x_T)$ for $\mathrm{P}_M(X_W \mid X_S \in \mathcal{S})$-a.a. $x_{\widehat{W}} = x_W \in \mathcal{X}_W$ and all $x_T \in \mathcal{X}_T$. Hence, we have

$$\begin{aligned} \mathrm{P}_{M_{|\mathcal{S}}}\left(\{X_{O\setminus T_i}(x_{T_i})\}_{1\le i\le n}\right) &= \left(\widehat{g}^{O\setminus T_1}(x_{T_1}, \cdot), \ldots, \widehat{g}^{O\setminus T_n}(x_{T_n}, \cdot)\right)_* \mathrm{P}_{M_{|\mathcal{S}}}(X_{\widehat{W}}) \\ &= \left(\widetilde{g}^{O\setminus T_1}(x_{T_1}, \cdot), \ldots, \widetilde{g}^{O\setminus T_n}(x_{T_n}, \cdot)\right)_* \mathrm{P}_M(X_W \mid X_S \in \mathcal{S}) \\ &= \left(g_{O\setminus T_1}^{V\setminus T_1}(x_{T_1}, \cdot), \ldots, g_{O\setminus T_n}^{V\setminus T_n}(x_{T_n}, \cdot)\right)_* \mathrm{P}_M(X_W \mid X_S \in \mathcal{S}) \\ &= \mathrm{P}_M\left(\{X_{O\setminus T_i}(x_{T_i})\}_{1\le i\le n} \mid X_S \in \mathcal{S}\right), \end{aligned}$$

since $g_{O\setminus T_i}^{V\setminus T_i}(x_W, x_{T_i}) = \widetilde{g}^{O\setminus T_i}(x_W, x_{T_i})$ for $\mathrm{P}_M(X_W)$-a.a. $x_W \in \mathcal{X}_W$ and all $x_{T_i} \in \mathcal{X}_{T_i}$ where $g^{V\setminus T_i} : \mathcal{X}_{T_i \cup S} \times \mathcal{X}_W \to \mathcal{X}_{V\setminus T_i}$ is the (essentially unique) solution function of $M$ w.r.t. $V \setminus T_i$ by Forré and Mooij (2025, Lemma 6.8.4). □

**Theorem 3.17** (Conditioning operation preserves as much causal information as possible)**.** *There are no mappings $(M, X_S \in \mathcal{S}) \mapsto \widetilde{M}$ that preserve more causal information than $(M, X_S \in \mathcal{S}) \mapsto M_{|X_S \in \mathcal{S}}$ in the following sense. Assume that the mapping $(M, X_S \in \mathcal{S}) \mapsto \widetilde{M}$ is such that for all $T_i \subseteq V$ and $x_{T_i} \in \mathcal{X}_{T_i}$*

$$\mathrm{P}_{\widetilde{M}}\big(\{X_{O\setminus T_i}(x_{T_i})\}_{1\le i\le n}\big) = \mathrm{P}_M\big(\{X_{O\setminus T_i}(x_{T_i})\}_{1\le i\le n} \mid X_S \in \mathcal{S}\big),$$

*and furthermore for some $T \subseteq \mathrm{Anc}_{G(M)}(S)$ and $x_T \in \mathcal{X}_T$*

$$\mathrm{P}_{\widetilde{M}}\left(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T)\right) = \mathrm{P}_M\left(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S}\right).$$

*Then it holds*

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

*Proof.* The proof is obvious by noting that

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

□

**Proposition 3.18** (Conditioning and marginalization commute)**.** *Assume Assumption 3.1 and let $L \subseteq V \setminus S$. Then we have $(M_{\setminus L})_{|X_S \in \mathcal{S}} \equiv (M_{|X_S \in \mathcal{S}})_{\setminus L}$.*

*Proof.* Write $(M_{|X_S \in \mathcal{S}})_{\setminus L} = (\widehat{V}, \widehat{W}, \widehat{\mathcal{X}}, \widehat{\mathrm{P}}, \widehat{f})$ and $(M_{\setminus L})_{|X_S \in \mathcal{S}} = (\overline{V}, \overline{W}, \overline{\mathcal{X}}, \overline{\mathrm{P}}, \overline{f})$.

First, it is easy to see that $\widehat{V} = \overline{V} = V \setminus (L \cup S)$. Let $\widetilde{g} : \mathcal{X}_W \to \mathcal{X}_{V \setminus L}$ be the (essentially unique) solution function of $M_{\setminus L}$. Then we have $g_S = \widetilde{g}_S$ $\mathrm{P}(X_W)$-a.s. Therefore, by the definition of $\widehat{W}$ and $\overline{W}$, we have $\widehat{W} = \overline{W}$ and $\widehat{\mathrm{P}}(X_{\widehat{W}}) = \overline{\mathrm{P}}(X_{\overline{W}}) = \mathrm{P}_M(X_W \mid X_S \in \mathcal{S})$. Furthermore, we have $\widehat{\mathcal{X}} = \overline{\mathcal{X}}$. By Bongers et al. (2021, Proposition 5.4), we have $\widehat{f}(x_{V \setminus (L \cup S)}, x_{\widehat{W}}) = \overline{f}(x_{V \setminus (L \cup S)}, x_{\overline{W}})$ for $\mathrm{P}_M(X_W \mid X_S \in \mathcal{S})$-a.a. $x_{\widehat{W}} = x_{\overline{W}} \in \mathcal{X}_{\widehat{W}}$ and all $x_{V \setminus (L \cup S)} \in \mathcal{X}_{V \setminus (L \cup S)}$. Overall, $(M_{\setminus L})_{|X_S \in \mathcal{S}} \equiv (M_{|X_S \in \mathcal{S}})_{\setminus L}$. □

**Proposition 3.19** (Conditioning and conditioning commute)**.** *Assume Assumption 3.1 with $S = S_1 \dot\cup S_2$ and $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$ where $\mathcal{S}_1 \subseteq \mathcal{X}_{S_1}$ and $\mathcal{S}_2 \subseteq \mathcal{X}_{S_2}$ are 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 and induce the same laws of potential outcomes. Also, $G(M_{|\mathcal{S}_1 \times \mathcal{S}_2})$ is a subgraph of $G((M_{|\mathcal{S}_1})_{|\mathcal{S}_2})$ and $G((M_{|\mathcal{S}_2})_{|\mathcal{S}_1})$. Furthermore, if*

*(i)* $\mathrm{Anc}_{G^a(M_{\setminus (V \setminus S_1)})}(S_1) \cap \mathrm{Anc}_{G^a(M_{\setminus (V \setminus S_2)})}(S_2) = \emptyset$*, or*

*(ii) we have*

$$\mathrm{P}\left(X_W \in \left(g_{S_1}^{-1}(\mathcal{S}_1) \triangle g_{S_2}^{-1}(\mathcal{S}_2)\right)\right) = 0,$$

*then $(M_{|\mathcal{S}_1})_{|\mathcal{S}_2} \equiv (M_{|\mathcal{S}_2})_{|\mathcal{S}_1} \equiv M_{|\mathcal{S}_1 \times \mathcal{S}_2}$.*

*Proof.* It is easy to see that $(M_{|\mathcal{S}_1})_{|\mathcal{S}_2}$, $(M_{|\mathcal{S}_2})_{|\mathcal{S}_1}$, and $M_{|\mathcal{S}_1 \times \mathcal{S}_2}$ are well defined given the assumption above. Write $O := V \setminus (S_1 \cup S_2)$. The counterfactual and potential-outcome equivalence among the three SCMs can be deduced by the following observation:

$$\begin{aligned}
\mathrm{P}_{M_{|\mathcal{S}_1 \times \mathcal{S}_2}}\left(\{X_O(x_{T_i})\}_{i=1}^n\right) &= \mathrm{P}_M\left(\{X_O(x_{T_i})\}_{i=1}^n \mid X_S \in \mathcal{S}_1 \times \mathcal{S}_2\right) \\
&= \mathrm{P}_{(M_{|\mathcal{S}_1})_{|\mathcal{S}_2}}\left(\{X_O(x_{T_i})\}_{i=1}^n\right) \\
&= \mathrm{P}_{(M_{|\mathcal{S}_2})_{|\mathcal{S}_1}}\left(\{X_O(x_{T_i})\}_{i=1}^n\right).
\end{aligned}$$

Now we show that $G(M_{|\mathcal{S}_1 \times \mathcal{S}_2})$ is a subgraph of $G((M_{|\mathcal{S}_1})_{|\mathcal{S}_2})$ and $G((M_{|\mathcal{S}_2})_{|\mathcal{S}_1})$. By Lemma 3.22, we can find a simple SCM $\widetilde{M}$ such that $\mathrm{Sib}_{G(\widetilde{M})}(\widetilde{S}_1 \cup \widetilde{S}_2) \cup \mathrm{Ch}_{G(\widetilde{M})}(\widetilde{S}_1 \cup \widetilde{S}_2) = \emptyset$ and $M_{|\mathcal{S}_1 \times \mathcal{S}_2} \equiv (\widetilde{M}_{|\widetilde{\mathcal{S}}_1 \times \widetilde{\mathcal{S}}_2})_{\setminus S_1 \cup S_2}$ and $(M_{|\mathcal{S}_1})_{|\mathcal{S}_2} \equiv (((\widetilde{M}_{|\widetilde{\mathcal{S}}_1})_{\setminus S_1})_{|\widetilde{\mathcal{S}}_2})_{\setminus S_2}$. We have by Proposition 3.18

$$(((\widetilde{M}_{|\widetilde{\mathcal{S}}_1})_{\setminus S_1})_{|\widetilde{\mathcal{S}}_2})_{\setminus S_2} \equiv (((\widetilde{M}_{|\widetilde{\mathcal{S}}_1})_{|\widetilde{\mathcal{S}}_2})_{\setminus S_1})_{\setminus S_2} \equiv ((\widetilde{M}_{|\widetilde{\mathcal{S}}_1})_{|\widetilde{\mathcal{S}}_2})_{\setminus S_1 \cup S_2}.$$

It suffices to show that $G((\widetilde{M}_{|\widetilde{\mathcal{S}}_1 \times \widetilde{\mathcal{S}}_2})_{\setminus S_1 \cup S_2})$ is a subgraph of $G(((\widetilde{M}_{|\widetilde{\mathcal{S}}_1})_{|\widetilde{\mathcal{S}}_2})_{\setminus S_1 \cup S_2})$. Write $M' := \widetilde{M}_{|\widetilde{\mathcal{S}}_1 \times \widetilde{\mathcal{S}}_2}$ and $M'' := (\widetilde{M}_{|\widetilde{\mathcal{S}}_1})_{|\widetilde{\mathcal{S}}_2}$. By definition, $M'$ and $M''$ have the same causal mechanisms and the same exogenous distribution, which is $\mathrm{P}_M(X_W \mid X_{S_1} \in \mathcal{S}_1, X_{S_2} \in \mathcal{S}_2)$. Therefore, we can use the same solution functions w.r.t. $S_1 \cup S_2$ for marginalizations. Hence, the directed edges in their graphs coincide. Note that the exogenous nodes of $\widetilde{M}_{|\widetilde{\mathcal{S}}_1 \times \widetilde{\mathcal{S}}_2}$ have the finest partition of $W$ given $X_{S_1} \in \mathcal{S}_1$ and $X_{S_2} \in \mathcal{S}_2$. Therefore, the bidirected edges of $\widetilde{M}_{|\widetilde{\mathcal{S}}_1 \times \widetilde{\mathcal{S}}_2}$ are a subset of those of $(\widetilde{M}_{|\widetilde{\mathcal{S}}_1})_{|\widetilde{\mathcal{S}}_2}$. This finishes the proof.

Finally, we show that, under the conditions of the proposition, we have $(M_{|\mathcal{S}_1})_{|\mathcal{S}_2} \equiv (M_{|\mathcal{S}_2})_{|\mathcal{S}_1} \equiv M_{|\mathcal{S}_1 \times \mathcal{S}_2}$. Write

$$\begin{aligned}(M_{|\mathcal{S}_1})_{|\mathcal{S}_2} &= (V^{12}, W^{12}, \mathcal{X}^{12}, f^{12}, \mathrm{P}^{12}) \\ (M_{|\mathcal{S}_2})_{|\mathcal{S}_1} &= (V^{21}, W^{21}, \mathcal{X}^{21}, f^{21}, \mathrm{P}^{21}) \\ M_{|\mathcal{S}_1 \times \mathcal{S}_2} &= (V^{1\times 2}, W^{1\times 2}, \mathcal{X}^{1\times 2}, f^{1\times 2}, \mathrm{P}^{1\times 2}).\end{aligned}$$

First, note that to establish $(M_{|\mathcal{S}_1})_{|\mathcal{S}_2} \equiv (M_{|\mathcal{S}_2})_{|\mathcal{S}_1} \equiv M_{|\mathcal{S}_1 \times \mathcal{S}_2}$, it suffices to show that $W^{12} = W^{21} = W^{1\times 2}$.[31] Indeed, it is easy to see that $V^{12} = V^{21} = V^{1\times 2}$. Given that $W^{12} = W^{21} = W^{1\times 2}$, we have $\mathcal{X}^{12} = \mathcal{X}^{21} = \mathcal{X}^{1\times 2}$ and $\mathrm{P}^{12} = \mathrm{P}^{21} = \mathrm{P}^{1\times 2}$. By the properties of marginalization, we have $f^{12} = f^{21} = f^{1\times 2}$ $\mathrm{P}^{12}$-a.s..

By the property of SCMs, there exist measurable functions $\widetilde{g}_{S_1} : \mathcal{X}_{\mathrm{Anc}_{G^a(M_{\setminus(V\setminus S_1)})}(S_1)\cap W} \to \mathcal{X}_{S_1}$ and $\widetilde{g}_{S_2} : \mathcal{X}_{\mathrm{Anc}_{G^a(M_{\setminus(V\setminus S_2)})}(S_2)\cap W} \to \mathcal{X}_{S_2}$ such that $g_{S_1}(x_W) = \widetilde{g}_{S_1}(x_{\mathrm{Anc}_{G^a(M_{\setminus(V\setminus S_1)})}(S_1)\cap W})$ and $g_{S_2}(x_W) = \widetilde{g}_{S_2}(x_{\mathrm{Anc}_{G^a(M_{\setminus(V\setminus S_2)})}(S_2)\cap W})$ for $\mathrm{P}(X_W)$-a.a. $x_W \in \mathcal{X}_W$. Since $\mathrm{Anc}_{G^a(M_{\setminus(V\setminus S_1)})}(S_1) \cap \mathrm{Anc}_{G^a(M_{\setminus(V\setminus S_2)})}(S_2) = \emptyset$, we have $W^{12} = W^{21}$. To see $W^{1\times 2} = W^{12}$, it suffices to note that $g^{-1}_{S_1\cup S_2}(\mathcal{S}_1 \times \mathcal{S}_2) = g^{-1}_{S_1}(\mathcal{S}_1) \cap g^{-1}_{S_2}(\mathcal{S}_2)$. If $\mathrm{P}(X_W \in (g^{-1}_{S_1}(\mathcal{S}_1) \triangle g^{-1}_{S_2}(\mathcal{S}_2))) = 0$, then we have $g^{-1}_{S_1}(\mathcal{S}_1) \overset{p}{=} g^{-1}_{S_2}(\mathcal{S}_2) \overset{p}{=} g^{-1}_{S_1}(\mathcal{S}_1) \cap g^{-1}_{S_2}(\mathcal{S}_2) = g^{-1}_{S_1\cup S_2}(\mathcal{S}_1 \times \mathcal{S}_2)$ where $\overset{p}{=}$ denotes equality up to a null set. This implies $W^{12} = W^{21} = W^{1\times 2}$.

□

**Lemma 3.22** (Conditioning on binary variable without children)**.** *Let $(M, X_S \in \mathcal{S})$ be a simple s-SCM and $(\widetilde{M}, X_{\widetilde{S}} = 1)$ be another simple s-SCM where $\widetilde{M} \coloneqq (\widetilde{V}, W, \widetilde{X}, \mathrm{P}, \widetilde{f})$ is such that $\widetilde{V} = V \dot\cup \{\widetilde{S}\}$, $\widetilde{\mathcal{X}} = \mathcal{X} \times \mathcal{X}_{\widetilde{S}} \coloneqq \mathcal{X} \times \{0, 1\}$ and $\widetilde{f}_{\widetilde{V}\setminus\widetilde{S}}(x_{\widetilde{V}}, x_W) = f_V(x_V, x_W)$ and $\widetilde{f}_{\widetilde{S}}(x_{\widetilde{V}}, x_W) = \mathbb{1}_{\mathcal{S}}(x_S)$. Then $M_{|X_S\in\mathcal{S}} \equiv (\widetilde{M}_{|X_{\widetilde{S}}=1})_{\setminus S}$.*

*Proof.* First note that $\widetilde{M}$ defined above is a simple SCM and $\mathrm{P}_{\widetilde{M}}(X_{\widetilde{S}} = 1) = \mathrm{P}_M(X_S \in \mathcal{S}) > 0$. We declare some notation. We write $M_{|X_S\in\mathcal{S}} = (\widehat{V}, \widehat{W}, \widehat{\mathcal{X}}, \widehat{\mathrm{P}}, \widehat{f})$ and $(\widetilde{M}_{|X_{\widetilde{S}}=1})_{\setminus S} = (\overline{V}, \overline{W}, \overline{\mathcal{X}}, \overline{\mathrm{P}}, \overline{f})$. It is easy to see that $\widehat{V} = \overline{V} = V \setminus S$ and $\widehat{\mathcal{X}} = \overline{\mathcal{X}} = \mathcal{X}_{V\setminus S}$. Also note that $g_S(\mathcal{S})^{-1} = \widetilde{g}_{\widetilde{S}}(1)^{-1}$ up to a P-null set where $g$ and $\widetilde{g}$ are the (essentially unique) solution functions of $M$ and $\widetilde{M}$ respectively. Then we can conclude that $\widehat{W} = \overline{W}$ and also $\widehat{\mathrm{P}} = \overline{\mathrm{P}} = \mathrm{P}_M(X_W \mid X_S \in \mathcal{S})$. It is also obvious that $\widehat{f}(x_{V\setminus S}, x_{\widehat{W}}) = \overline{f}(x_{V\setminus S}, x_{\overline{W}}) = f_{V\setminus S}(x_{V\setminus S}, g^S(X_{V\setminus S}, x_W), x_W)$ where $g^S$ is the (essentially unique) solution function of $M$ w.r.t. $S$, for $\mathrm{P}_M(X_W \mid X_S \in \mathcal{S})$-a.a. $x_{\overline{W}} = x_{\widehat{W}} = x_W \in \mathcal{X}_W$ and all $x_{V\setminus S} \in \mathcal{X}_{V\setminus S}$. Overall, we have $M_{|X_S\in\mathcal{S}} \equiv (\widetilde{M}_{|X_{\widetilde{S}}=1})_{\setminus S}$. □

**Theorem 3.25** (Main result II: Graphical separation in conditioning graph)**.** *Let $G = (V, E^d, E^b)$ be a DMG and $S \subseteq V$ a set of nodes. Then for any subsets of nodes $A, B, C \subseteq V$ such that*

$$S \cap (A \cup B \cup C) = \emptyset,$$

*it holds that*

$$A \overset{\sigma/d}{\underset{G_{|S}}{\perp}} B \mid C \quad \Longrightarrow \quad A \overset{\sigma/d}{\underset{G}{\perp}} B \mid C \cup S.$$

[31] Strictly speaking they are not exactly equal to each other but are isomorphic. For simplicity, we see being isomorphic as equal.

*If furthermore $S$ is a singleton set with $\mathrm{Ch}_G(S) = \emptyset$ and $C \cap \mathrm{Anc}_G(S) = \emptyset$, then we have*

$$A \overset{\sigma/d}{\underset{G}{\perp}} B \mid C \cup S \quad \Longrightarrow \quad A \overset{\sigma/d}{\underset{G_{|S}}{\perp}} B \mid C.$$

*Proof.* We show the first statement. By Proposition 3.27, we can assume WLOG that $S = \{s\}$ is a singleton set. To show

$$A \overset{\sigma}{\underset{G_{|S}}{\perp}} B \mid C \quad \Longrightarrow \quad A \overset{\sigma}{\underset{G}{\perp}} B \mid C \cup S,$$

it suffices to show

$$A \overset{\sigma}{\underset{G}{\not\perp}} B \mid C \cup S \quad \Longrightarrow \quad A \overset{\sigma}{\underset{G_{|S}}{\not\perp}} B \mid C.$$

Let $\pi : v_0 \ast\!\!\rightarrow \cdots \ast\!\!\rightarrow v_n$ be a $\sigma$-open walk between $A$ and $B$ given $C \cup S$ in $G$ such that all the colliders are in $C \cup S$ (Forré and Mooij, 2025, Proposition 3.3.6). WLOG we can assume that $s$ appears on $\pi$ at most once. Indeed, assume that $\pi$ is of the form $v_0 \ast\!\!\rightarrow \cdots \ast\!\!\rightarrow v_i \ast\!\!\rightarrow s \ast\!\!\rightarrow \cdots \ast\!\!\rightarrow s \ast\!\!\rightarrow v_j \ast\!\!\rightarrow \cdots v_n$ and the subwalks $\pi(v_0, v_i)$ and $\pi(v_j, v_n)$ do not contain $s$. If we have $v_i \ast\!\!\rightarrow s$ and $s \leftarrow\!\!\ast\, v_j$, then the walk $\pi(v_0, v_i) \oplus \pi(v_i, s) \oplus \pi(s, v_j) \oplus \pi(v_j, n)$ is $\sigma$-open given $C \cup S$. If we have $v_i \leftarrow s$ and $s \leftarrow\!\!\ast\, v_j$, then $v_i$ must be in the same strongly connected component of $s$ since otherwise $\pi$ is blocked by $S$. Therefore, the walk $\pi(v_0, v_i) \oplus \pi(v_i, s) \oplus \pi(s, v_j) \oplus \pi(v_j, n)$ is $\sigma$-open given $C \cup S$. The same conclusion holds if $v_i \ast\!\!\rightarrow s$ and $s \rightarrow v_j$, and likewise if $v_i \leftarrow s$ and $s \rightarrow v_j$. If $s$ occurs on $\pi$ as a non-collider, i.e., $v_{i-1} \ast\!\!\rightarrow s \rightarrow v_{i+1}$ or $v_{i-1} \leftarrow s \rightarrow v_{i+1}$ or $v_{i-1} \leftarrow s \leftarrow\!\!\ast\, v_{i+1}$, then, by replacing the segments by $v_{i-1} \ast\!\!\rightarrow v_{i+1}$ or $v_{i-1} \leftrightarrow v_{i+1}$ or $v_{i-1} \leftarrow\!\!\ast\, v_{i+1}$ respectively and keeping other parts of $\pi$ intact, we have a $\sigma$-open walk from $A$ to $B$ given $C$ in $G_{|S}$. So consider the case where $s$ does not occur on $\pi$ as a non-collider in the following. If all colliders on $\pi$ are in $C$ and $\pi$ does not contain $s$, then $\pi$ is $\sigma$-open between $A$ and $B$ given $C$ in $G_{|S}$. We consider the case where $s$ occurs on $\pi$ as collider. Define $v_l$ and $v_r$ to be the most left and most right nodes on $\pi$ respectively that are in $\mathrm{Anc}_G(S) \cup \mathrm{Sib}_G(\mathrm{Anc}_G(S))$. So $\pi$ is of the form $v_0 \ast\!\!\rightarrow \cdots \ast\!\!\rightarrow v_l \ast\!\!\rightarrow \cdots \rightarrow s \leftarrow \cdots \leftarrow\!\!\ast\, v_r \ast\!\!\rightarrow \cdots \ast\!\!\rightarrow v_n$. We replace the subwalk $\pi(v_l, v_r)$ with $v_l \leftrightarrow v_r$ on $\pi$ to construct a new walk $\widetilde{\pi}$. Note that $v_l$ cannot become a collider on $\widetilde{\pi}$ if it is a non-collider on $\pi$ and similarly for $v_r$. If $v_l$ is a collider on $\widetilde{\pi}$, it must be in $C$. Similarly for $v_r$. If $v_l$ is a non-collider on $\pi$, it must be unblockable or not in $C$. If it is unblockable on $\pi$, it remains so on $\widetilde{\pi}$. Similarly for $v_r$. Hence, $\widetilde{\pi}$ is $\sigma$-open in $G_{|S}$ from $A$ to $B$ given $C$.

Overall, given any $\sigma$-open walk between $A$ and $B$ given $C \cup S$ in $G$, we can find a $\sigma$-open walk between $A$ and $B$ given $C$ in $G_{|S}$. This shows that

$$A \overset{\sigma}{\underset{G}{\not\perp}} B \mid C \cup S \quad \Longrightarrow \quad A \overset{\sigma}{\underset{G_{|S}}{\not\perp}} B \mid C.$$

We now show

$$A \overset{\sigma}{\underset{G}{\perp}} B \mid C \cup S \quad \Longrightarrow \quad A \overset{\sigma}{\underset{G_{|S}}{\perp}} B \mid C$$

under the conditions that $C \cap \mathrm{Anc}_G(S) = \emptyset$, $\mathrm{Ch}_G(S) = \emptyset$, and $S$ is a singleton set. Assume that $\pi$ is a $\sigma$-open walk between $A$ and $B$ given $C$ in $G_{|S}$ such that all the colliders are in $C$. We shall construct a $\sigma$-open walk between $A$ and $B$ given $C \cup S$ in $G$.

For all edges in $\pi$ that are also in $G$, we keep them untouched. Note that since $\mathrm{Ch}_G(S) = \emptyset$, there do not exist directed edges in $G_{|S}$ that are not in $G$. Therefore, if all bidirected edges on $\pi$ in $G_{|S}$ are also in $G$, then we are done. Thus, we are left with the case where there are some bidirected edges $v_i \dashleftarrow\dashrightarrow v_{i+1}$ of $\pi$ not in $G$. Note that since $S$ is singleton and the possibilities of $v_i \dashleftarrow\dashrightarrow s \rightarrow v_{i+1}$ and $v_i \leftarrow s \dashleftarrow\dashrightarrow v_{i+1}$ and $v_i \leftarrow s \rightarrow v_{i+1}$ are excluded by the assumption $\mathrm{Ch}_G(S) = \emptyset$, we can replace those bidirected edges $v_i \dashleftarrow\dashrightarrow v_{i+1}$ with $v_i \dashleftarrow\!\!*\!\!\dashrightarrow w_1 \rightarrow \cdots \rightarrow w_k \rightarrow s \leftarrow w_{k+1} \leftarrow \cdots \leftarrow w_m \dashleftarrow\!\!*\!\!\dashrightarrow v_{i+1}$ ($k$ could be zero and $m$ could be $k$) and get a new walk $\widetilde{\pi}$ in $G$. Next, we show that $\widetilde{\pi}$ cannot be blocked by $C \cup S$ at $v_i$ or $v_{i+1}$ in $G$.

If $v_i$ or $v_{i+1}$ is an endnode of $\pi$, then $\widetilde{\pi}$ cannot be blocked at that node in $G$. Thus, we assume that $v_i$ and $v_{i+1}$ are not endnodes of $\pi$. Suppose that in $\pi$ we have $\cdots \dashleftarrow\!\!*\!\!\dashrightarrow v_i \dashleftarrow\dashrightarrow v_{i+1} \cdots$. Since $v_i$ is a collider on $\pi$, by the assumption on $\pi$, we have $v_i \in C$. Then it is impossible to have $\cdots \dashleftarrow\!\!*\!\!\dashrightarrow v_i \rightarrow w_1 \rightarrow \cdots \rightarrow w_k \rightarrow s$ in $G$, since in this case we would have $\mathrm{Anc}_G(S) \cap C \neq \emptyset$. Hence, we must have $\cdots \dashleftarrow\!\!*\!\!\dashrightarrow v_i \dashleftarrow\dashrightarrow w_1 \rightarrow \cdots w_k \rightarrow s$, and then we know that this walk is $\sigma$-open given $C$ at $v_i$ in $G$.

Now we suppose $\cdots \leftarrow v_i \dashleftarrow\dashrightarrow v_{i+1}$. We have to show that $\cdots \leftarrow v_i \dashleftarrow\!\!*\!\!\dashrightarrow w_1 \rightarrow \cdots \rightarrow w_k \rightarrow s$ is $\sigma$-open given $C \cup S$ at $v_i$. To check it, first assume $\cdots \leftarrow v_i \dashleftarrow\dashrightarrow w_1 \rightarrow \cdots \rightarrow w_k \rightarrow s$. In this case, if $v_i \notin C$, then this is obviously $\sigma$-open. If $v_i \in C$, then it is also $\sigma$-open, since $v_i$ must point to the same strongly connected component. Otherwise, $\pi$ cannot be $\sigma$-open given $C$ at $v_i$ in $G_{|S}$. Second, we assume $\cdots \leftarrow v_i \rightarrow w_1 \rightarrow \cdots \rightarrow w_k \rightarrow s$. In this case $v_i \notin C$, since if $v_i \in C$ then $\mathrm{Anc}(C) \cap S \neq \emptyset$. Thus, $\widetilde{\pi}$ is $(C \cup S)$-$\sigma$-open at $v_i$ in $G$.

A similar argument can be made for $v_{i+1}$. Then we have constructed a $\sigma$-open walk between $A$ and $B$ given $C \cup S$ in $G$. This contradicts the fact that $A \overset{\sigma}{\underset{G}{\perp}} B \mid C \cup S$. Therefore, there is no $\sigma$-open walk between $A$ and $B$ given $C$ in $G_{|S}$. So we have $A \overset{\sigma}{\underset{G_{|S}}{\perp}} B \mid C$.

Almost the same argument also applies to the case of $d$-separation. □

**Proposition 3.27** (Graph conditioning commutes with marginalization, conditioning and intervention)**.** *Let $G = (V, E^d, E^b)$ be a DMG.*

*(1) Let $L \subseteq V$ and $S \subseteq V$ be two disjoint subsets of nodes from $G$. Then we have*

$$\left(G_{\backslash L}\right)_{|S} = \left(G_{|S}\right)_{\backslash L}.$$

*(2) Let $S_1, S_2 \subseteq V$ be two disjoint subsets. Then we have*

$$\left(G_{|S_1}\right)_{|S_2} = \left(G_{|S_2}\right)_{|S_1} \subseteq G_{|S_1 \cup S_2}.$$

*(3) Let $T \subseteq V$ and $S \subseteq V$ be two disjoint subsets of nodes from $G$ such that $T \cap \mathrm{Anc}_G(S) = \emptyset$. Then we have*

$$\left(G_{\mathrm{do}(T)}\right)_{|S} = \left(G_{|S}\right)_{\mathrm{do}(T)}.$$

*Proof.* We show the first two results. Denoting by $G_{\mathrm{abe}(S)}$ the graph obtained by the first step of Definition 3.23 allows us to write $G_{|S} = (G_{\mathrm{abe}(S)})_{\backslash S}$. By Lemma C.1 and the fact that marginalizations of two disjoint sets commute (Bongers et al., 2021, Proposition 5.8), we have

$$(G_{\backslash L})_{|S} = ((G_{\backslash L})_{\mathrm{abe}(S)})_{\backslash S} = ((G_{\mathrm{abe}(S)})_{\backslash L})_{\backslash S} = ((G_{\mathrm{abe}(S)})_{\backslash S})_{\backslash L} = (G_{|S})_{\backslash L},$$

and by Lemma C.2

$$\begin{aligned}\left(G_{|S_1}\right)_{|S_2} &= \left(\left((G_{\mathrm{abe}(S_1)})_{\setminus S_1}\right)_{\mathrm{abe}(S_2)}\right)_{\setminus S_2} \\ &= \left(\left((G_{\mathrm{abe}(S_1)})_{\mathrm{abe}(S_2)}\right)_{\setminus S_2}\right)_{\setminus S_1} \\ &= \left(\left((G_{\mathrm{abe}(S_2)})_{\setminus S_2}\right)_{\mathrm{abe}(S_1)}\right)_{\setminus S_1} = (G_{|S_2})_{|S_1}.\end{aligned}$$

Since $(G_{\mathrm{abe}(S_1)})_{\mathrm{abe}(S_2)} \subseteq G_{\mathrm{abe}(S_1 \cup S_2)}$, we have $\left(G_{|S_1}\right)_{|S_2} \subseteq G_{|S_1 \cup S_2}$.

We now show the third result. It suffices to consider $T = \{t\}$ and show that $(G_{\mathrm{do}(T)})_{\mathrm{abe}(S)} = (G_{\mathrm{abe}(S)})_{\mathrm{do}(T)}$. Note that there are two cases. One is $t \notin \mathrm{Sib}_G(\mathrm{Anc}_G(S))$ and the other one is $t \in \mathrm{Sib}_G(\mathrm{Anc}_G(S)) \setminus \mathrm{Anc}_G(S)$. Local configuration of a node means the edges adjacent to that node in the graph. In the first case, the local configuration of $t$ is independent of performing $\mathrm{abe}(S)$. In the second case, first intervening on $t$ breaks the bidirected edge between $t$ and $\mathrm{Anc}_G(S)$ and second performing $\mathrm{abe}(S)$ keeps the local configuration of $t$ untouched, while first performing $\mathrm{abe}(S)$ adds some bidirected edges between $t$ and nodes in $\mathrm{Anc}_G(S) \cup \mathrm{Sib}_G(\mathrm{Anc}_G(S))$, but second intervening on $t$ then throws these bidirected edges away. Hence, no matter what order of the conditioning and the intervening are applied, one still ends up with the same graph. □

**Proposition 3.29** (Main result III: DMG conditioning is compatible with SCM conditioning)**.** *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 $G(M)_{|S}$. If furthermore $S = \{s_1, \ldots, s_n\}$ and $\mathcal{S} = \bigtimes_{i=1}^n \mathcal{S}_i$ with $\mathcal{S}_i \subseteq \mathcal{X}_{s_i}$ measurable for $i = 1, \ldots, n$, then $G(M_{|X_S \in \mathcal{S}})$ is a subgraph of $((G(M)_{|s_1})_{\ldots})_{|s_n}$.*

*Proof.* We call a subset $A$ of $V$ ancestral if $\mathrm{Anc}_{G(M)}(A) = A$ and call an SCM $M$ ancestrally uniquely solvable if for every ancestral subset $A$ of $V$ the SCM $M$ is essentially uniquely solvable w.r.t. $A$. Since simple SCMs are ancestrally uniquely solvable, we have that $G(M_{\setminus S})$ is a subgraph of $G(M)_{\setminus S}$ by Bongers et al. (2021, Proposition 5.11). Let $g : \mathcal{X}_W \to \mathcal{X}_V$ be the (essentially) unique solution function of $M$. Note that there exists a measurable map $\widetilde{g} : \mathcal{X}_{W \cap \mathrm{Anc}_{G^a}(S)} \to \mathcal{X}_S$ such that $\widetilde{g} = g_S$ $\mathrm{P}(X_W)$-a.s. So, exogenous variables that are not in $W \cap \mathrm{Anc}_{G^a}(S)$ will not merge. Denote by $M_{\mathrm{merge}}$ the SCM obtained from the merging operation. This implies that $G(M_{\mathrm{merge}})$ is a subgraph of $G(M)_{\mathrm{abe}(S)}$. Since $\mathrm{P}_M(X_W \mid X_S \in \mathcal{S}) \ll \mathrm{P}(X_W)$, $G((M_{\mathrm{merge}})_{\mathrm{update}})$ is a subgraph of $G(M_{\mathrm{merge}})$, where $M_{\mathrm{update}}$ denotes the SCM obtained from $M$ via updating the exogenous distribution to the posterior given $X_S \in \mathcal{S}$. Definition of the conditioned DMG and Bongers et al. (2021, Proposition 5.11) yield

$$\begin{aligned}G(M_{|X_S \in \mathcal{S}}) &= G(((M_{\mathrm{merge}})_{\mathrm{update}})_{\setminus S}) \\ &\subseteq G((M_{\mathrm{merge}})_{\mathrm{update}})_{\setminus S} \subseteq G(M_{\mathrm{merge}})_{\setminus S} \subseteq (G(M)_{\mathrm{abe}(S)})_{\setminus S} = G(M)_{|S},\end{aligned}$$

since $(M_{\mathrm{merge}})_{\mathrm{update}}$ is simple.

We now show the second claim. From the last part, we have that $G(M_{|X_{s_1} \in \mathcal{S}_1})$ is a subgraph of $G(M)_{|s_1}$. Note that if $G_1$ is a subgraph of $G_2$ then $(G_1)_{|S}$ is a subgraph of $(G_2)_{|S}$. By Proposition 3.19, it holds that $G(M_{|X_{\{s_1,s_2\}} \in \mathcal{S}_1 \times \mathcal{S}_2})$ is a subgraph of $G((M_{|X_{s_1} \in \mathcal{S}_1})_{|X_{s_2} \in \mathcal{S}_2})$ and is therefore a subgraph of $(G(M)_{|s_1})_{|s_2}$. Hence, we can conclude that $G(M_{|X_S \in \mathcal{S}})$ is a subgraph of $((G(M)_{|s_1})_{\ldots})_{|s_n}$. □

**Lemma 3.30.** *Let $X_A, X_B, X_C$ and $X_S$ be random variables defined on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and $X_S$ take values in a standard measurable space $(\mathcal{X}_S, \mathcal{B}_{\mathcal{X}_S})$. Then the first statement implies the second statement:*

*(1) $X_A \perp\!\!\!\perp X_B \mid X_C, X_S \in \mathcal{H}$ for all $\mathcal{H} \in \mathcal{B}_{\mathcal{X}_S}$ with positive probability.*

*(2) $X_A \perp\!\!\!\perp X_B \mid X_C, X_S$.*

*Proof.* Let $\{\mathcal{H}_n\}_{n=1}^{\infty}$ be a countable generator of $\mathcal{B}_{\mathcal{X}_S}$ such that $\mathrm{P}(X_S \in \mathcal{H}_n) > 0$ for all $n$, which exists since $(\mathcal{X}_S, \mathcal{B}_{\mathcal{X}_S})$ is standard. Define

$$\begin{aligned} \mathcal{G}_n &\coloneqq \sigma(X_C) \vee \sigma(\mathbb{1}(X_S \in \mathcal{H}_1), \ldots, \mathbb{1}(X_S \in \mathcal{H}_n)) \\ \text{and } \mathcal{G}_\infty &\coloneqq \sigma(X_C) \vee \sigma(\mathbb{1}(X_S \in \mathcal{H}_1), \mathbb{1}(X_S \in \mathcal{H}_2), \ldots). \end{aligned}$$

We suppose that $X_A \perp\!\!\!\perp X_B \mid X_C, X_S \in \mathcal{H}$ for all $\mathcal{H} \in \mathcal{B}_{\mathcal{X}_S}$ such that $\mathrm{P}(X_S \in \mathcal{H}) > 0$. Then we have for all measurable subsets $\mathcal{C} \subseteq \mathcal{X}_A$ and $\mathcal{D} \subseteq \mathcal{X}_B$ and for all $n \in \mathbb{N}$

$$\mathrm{P}\left((X_A, X_B) \in \mathcal{C} \times \mathcal{D} \mid \mathcal{G}_n\right) = \mathrm{P}(X_A \in \mathcal{C} \mid \mathcal{G}_n) \cdot \mathrm{P}(X_B \in \mathcal{D} \mid \mathcal{G}_n) \quad \text{P-a.s.}$$

Lévy's upward theorem (Williams, 1991) implies that $\mathrm{P}(\ldots \mid \mathcal{G}_n) \overset{\text{a.s.}}{\to} \mathrm{P}(\ldots \mid \mathcal{G}_\infty) = \mathrm{P}(\ldots \mid X_C, X_S)$ as $n \to \infty$. Therefore,

$$\mathrm{P}((X_A, X_B) \in \mathcal{C} \times \mathcal{D} \mid X_C, X_S) = \mathrm{P}(X_A \in \mathcal{C} \mid X_C, X_S) \cdot \mathrm{P}(X_B \in \mathcal{D} \mid X_C, X_S) \quad \text{P-a.s.},$$

which means that $X_A \perp\!\!\!\perp X_B \mid X_C, X_S$. □

**Lemma C.1** (The first step of Definition 3.23 commutes with marginalization)**.** *Let $G = (V, E^d, E^b)$ be a DMG and $S, L \subseteq V$ be two disjoint subsets. Then we have $(G_{\mathrm{abe}(S)})_{\backslash L} = (G_{\backslash L})_{\mathrm{abe}(S)}$.*

*Proof.* The two graphs $(G_{\mathrm{abe}(S)})_{\backslash L}$ and $(G_{\backslash L})_{\mathrm{abe}(S)}$ have the same set of nodes. We show that they have the same set of edges. It is easy to see that $v \dashrightarrow u$ is in $(G_{\mathrm{abe}(S)})_{\backslash L}$ iff $v \dashrightarrow u$ is in $(G_{\backslash L})_{\mathrm{abe}(S)}$. We show that for any $v, u \in V \setminus L$, we have $v \dashleftarrow\dashrightarrow u$ in $(G_{\mathrm{abe}(S)})_{\backslash L}$ iff $v \dashleftarrow\dashrightarrow u$ is in $(G_{\backslash L})_{\mathrm{abe}(S)}$.

Suppose that $v \dashleftarrow\dashrightarrow u$ is in $(G_{\mathrm{abe}(S)})_{\backslash L}$. There are two possibilities: (i) $v \dashleftarrow\dashrightarrow u$ is in $G_{\mathrm{abe}(S)}$ and (ii) $v \dashleftarrow\dashrightarrow u$ is not in $G_{\mathrm{abe}(S)}$. We now consider the first case. We will show that if $v \dashleftarrow\dashrightarrow u$ is in $G$, then it must be in $(G_{\backslash L})_{\mathrm{abe}(S)}$. By the properties of graphical marginalization, we have $\mathrm{Anc}_{G_{\backslash L}}(S) = \mathrm{Anc}_G(S) \setminus L$ and $\mathrm{Sib}_G(\mathrm{Anc}_G(S)) \setminus L \subseteq \mathrm{Sib}_{G\backslash L}(\mathrm{Anc}_{G_{\backslash L}}(S))$. This implies $\mathrm{SA}_G(S) \setminus L \subseteq \mathrm{SA}_{G_{\backslash L}}(S)$, where $\mathrm{SA}_G(S) \coloneqq \mathrm{Anc}_G(S) \cup \mathrm{Sib}_G(\mathrm{Anc}_G(S))$. So, if $u, v \in \mathrm{SA}_G(S)$, then $v \dashleftarrow\dashrightarrow u$ is in $(G_{\backslash L})_{\mathrm{abe}(S)}$. We now consider the second case. The bidirected edge $v \dashleftarrow\dashrightarrow u$ must come from some bifurcation in $G_{\mathrm{abe}(S)}$. If the bifurcation consists of only edges in $G$, then $v \dashleftarrow\dashrightarrow u$ must also occur in $(G_{\backslash L})_{\mathrm{abe}(S)}$. If the bifurcation contains edges not in $G$, then it must be of the form $v \dashleftarrow l_1 \dashleftarrow \cdots \dashleftarrow l_{k-1} \dashleftarrow\dashrightarrow l_k \dashrightarrow \cdots \dashrightarrow l_{n-1} \dashrightarrow u$ with $l_{k-1}, l_k \in \mathrm{SA}_G(S)$, where $l_i \in L$ for all $i$. This forces $v, u \in \mathrm{SA}_{G_{\backslash L}}(S)$. Therefore, $v \dashleftarrow\dashrightarrow u$ is in $(G_{\backslash L})_{\mathrm{abe}(S)}$.

Suppose that $v \dashleftarrow\dashrightarrow u$ is in $(G_{\backslash L})_{\mathrm{abe}(S)}$. If there is a bifurcation in $G$ between $v$ and $u$ with all intermediate nodes in $L$, then the same bifurcation exists in $G_{\mathrm{abe}(S)}$. Therefore, marginalizing out $L$ in $G_{\mathrm{abe}(S)}$ creates $v \dashleftarrow\dashrightarrow u$ in $(G_{\mathrm{abe}(S)})_{\backslash L}$. If $v \dashleftarrow\dashrightarrow u$ comes from the "abe" operation on $G_{\backslash L}$, then $v, u \in \mathrm{SA}_{G_{\backslash L}}(S)$. If $v, u \in \mathrm{Anc}_{G_{\backslash L}}(S)$, then $v, u \in \mathrm{Anc}_G(S)$. Assume

$v \in \mathrm{Sib}_{G_{\backslash L}}(\mathrm{Anc}_{G_{\backslash L}}(S)) \setminus \mathrm{Anc}_{G_{\backslash L}}(S)$. Then $v \dashleftarrow\dashrightarrow \widetilde{v}$ must be in $G_{\backslash L}$ for some $\widetilde{v} \in \mathrm{Anc}_{G_{\backslash L}}(S)$. We also have $u \dashleftarrow\dashrightarrow \widetilde{u}$ in $G_{\backslash L}$ for some $\widetilde{u} \in \mathrm{Anc}_{G_{\backslash L}}(S)$ if $u \notin \mathrm{Anc}_{G_{\backslash L}}(S)$. If these bidirected edges are present in $G$, then we are done. So we assume that $v \dashleftarrow\dashrightarrow \widetilde{v}$ is not present in $G$. Then there must exist bifurcations of the form $v \leftarrow l_1 \leftarrow \cdots \leftarrow l_{k-1} \dashleftarrow\dashrightarrow l_k \rightarrow \cdots \rightarrow l_{n-1} \rightarrow \widetilde{v}$ with $l_i \in L$ and $\widetilde{u} \leftarrow \widetilde{l}_1 \leftarrow \cdots \leftarrow \widetilde{l}_{j-1} \dashleftarrow\dashrightarrow \widetilde{l}_j \rightarrow \cdots \rightarrow \widetilde{l}_{m-1} \rightarrow u$ with $\widetilde{l}_i \in L$ ($m = 1$ if $u \dashleftarrow\dashrightarrow \widetilde{u}$ is present in $G$ or $u = \widetilde{u}$) in $G$. Then $l_{k-1}, \widetilde{l}_j \in \mathrm{SA}_G(S)$. Therefore, we have $v \leftarrow l_1 \leftarrow \cdots \leftarrow l_{k-1} \dashleftarrow\dashrightarrow \widetilde{l}_j \rightarrow \cdots \rightarrow \widetilde{l}_{m-1} \rightarrow u$ or $v \leftarrow l_1 \leftarrow \cdots \leftarrow l_{k-1} \dashleftarrow\dashrightarrow u$ in $G_{\mathrm{abe}(S)}$. Finally, $v \dashleftarrow\dashrightarrow u$ is in $(G_{\mathrm{abe}(S)})_{\backslash L}$. This finishes the proof. □

**Lemma C.2** (The first step of Definition 3.23 commutes with itself)**.** *Let $G = (V, E^d, E^b)$ be a DMG and $S_1, S_2 \subseteq V$ be two disjoint subsets. Then we have*

$$(G_{\mathrm{abe}(S_1)})_{\mathrm{abe}(S_2)} = (G_{\mathrm{abe}(S_2)})_{\mathrm{abe}(S_1)} \subseteq G_{\mathrm{abe}(S_1 \cup S_2)}.$$

*Proof.* To simplify the notation, we write $G_{ij} \coloneqq (G_{\mathrm{abe}(S_i)})_{\mathrm{abe}(S_j)}$. Write $D \coloneqq (\mathrm{SA}_G(S_1) \cap \mathrm{SA}_G(S_2)) \cap (\mathrm{Anc}_G(S_1) \cup \mathrm{Anc}_G(S_2))$. There are two cases: (i) $D = \emptyset$, and (ii) $D \neq \emptyset$. In the first case, $\mathrm{SA}_G(S_1) \cap \mathrm{SA}_G(S_2) = \emptyset$ or $\mathrm{SA}_G(S_1) \cap \mathrm{SA}_G(S_2)$ does not contain ancestors of $S_1 \cup S_2$ in $G$. Then it is not hard to see that adding bidirected edges to nodes in $\mathrm{SA}_G(S_1)$ and adding bidirect edges to nodes in $\mathrm{SA}_G(S_2)$ are independent. Therefore, we have $a \dashleftarrow\dashrightarrow b$ for every $a, b \in \mathrm{SA}_G(S_1)$ and $c \dashleftarrow\dashrightarrow d$ for every $c, d \in \mathrm{SA}_G(S_2)$ in both $G_{12}$ and $G_{21}$, and there are no other bidirected edges added to $G$. Hence, $G_{12} = G_{21}$ in this case. We now consider the second case. Pick an arbitrary node $a \in D$. By the definition of the set $D$, it holds $a \in \mathrm{Anc}_G(S_1) \cup \mathrm{Anc}_G(S_2)$. If $a \in \mathrm{Anc}_G(S_1) \cap \mathrm{Anc}_G(S_2)$, then we have $b \dashleftarrow\dashrightarrow c$ for all $b, c \in \mathrm{SA}_G(S_1) \cup \mathrm{SA}_G(S_2)$ in both $G_{12}$ and $G_{21}$, and there are no other bidirected edges added to $G$. In the following, WLOG, we assume $a \in \mathrm{Anc}_G(S_1)$ but $a \notin \mathrm{Anc}_G(S_2)$ by the symmetry of $\mathrm{Anc}_G(S_1)$ and $\mathrm{Anc}_G(S_2)$. Note that we have $a \dashleftarrow\dashrightarrow \widetilde{a}$ for some $\widetilde{a} \in \mathrm{Anc}_G(S_2)$. Then we have $\widetilde{a} \dashleftarrow\dashrightarrow d$ for all $d \in \mathrm{SA}_G(S_1)$ in $G_1$. This implies $\mathrm{SA}_G(S_1) \subseteq \mathrm{SA}_{G_1}(S_2)$ and hence $\mathrm{SA}_{G_1}(S_2) = \mathrm{SA}_G(S_1) \cup \mathrm{SA}_G(S_2)$. Overall, in $G_{12}$ we have bidirected edges between any two nodes in $\mathrm{SA}_G(S_1) \cup \mathrm{SA}_G(S_2)$ added to $G$ and no other bidirected edges are added. By the symmetry of $S_1$ and $S_2$, we have the same argument and conclusion hold for $G_{21}$. This concludes that $G_{12} = G_{21}$.

We now show the second claim. For that, observe that $G_{\mathrm{abe}(S_1)} \subseteq G_{\mathrm{abe}(S_1 \cup S_2)}$ and $G_{\mathrm{abe}(S_1 \cup S_2)} = (G_{\mathrm{abe}(S_1 \cup S_2)})_{\mathrm{abe}(S_2)}$. □

**Lemma C.3.** *Let $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ be a simple SCM. Let $g : \mathcal{X}_W \to \mathcal{X}_V$ be the (essentially unique) solution function of $M$. Let $A \subseteq V$ and write $B \coloneqq V \setminus A$. Let $g^B : \mathcal{X}_A \times \mathcal{X}_W \to \mathcal{X}_B$ be the solution function of $M$ w.r.t. $B$. Then*

$$g_B(x_W) = g^B(g_A(x_W), x_W)$$

*for $\mathrm{P}(X_W)$-a.a. $x_W \in \mathcal{X}_W$.*

*Proof.* For $\mathrm{P}(X_W)$-a.a. $x_W \in \mathcal{X}_W$ and all $x_V \in \mathcal{X}_V$, we have

$$\begin{cases} x_A = f_A(x_V, x_W) \\ x_B = f_B(x_V, x_W) \end{cases} \iff \begin{cases} x_A = g_A(x_W) \\ x_B = g_B(x_W). \end{cases}$$

Since $g^B$ is the essentially unique solution function of $M$ w.r.t. $B$, for $\mathrm{P}(X_W)$-a.a. $x_W \in \mathcal{X}_W$ and all $x_V \in \mathcal{X}_V$ we have

$$x_B = f_B(x_V, x_W) \Longleftrightarrow x_B = g^B(x_A, x_W).$$

Hence, for $\mathrm{P}(X_W)$-a.a. $x_W \in \mathcal{X}_W$ and all $x_V \in \mathcal{X}_V$ we have

$$\begin{cases} x_A = f_A(x_V, x_W) \\ x_B = f_B(x_V, x_W) \end{cases} \Longleftrightarrow \begin{cases} x_A = g_A(x_W) \\ x_B = g^B(x_A, x_W) \end{cases} \Longleftrightarrow \begin{cases} x_A = g_A(x_W) \\ x_B = g^B(g_A(x_W), x_W). \end{cases}$$

Since $M$ is uniquely solvable, we can conclude for $\mathrm{P}(X_W)$-a.a. $x_W \in \mathcal{X}_W$

$$g_B(x_W) = g^B(g_A(x_W), x_W).$$

□

# D Discussions on conditioning operation for SCMs

## D.1 SCMs are not flexible enough for representing s-SCMs

In this subsection, we show that, in general, it is impossible to find a simple SCM encoding all the causal semantics of a simple s-SCM. This gives an answer to Question Q1: the causal semantics of the ancestors of selection nodes cannot be preserved in general under the framework of simple SCMs and Theorem 3.14 shows that the causal semantics of the non-ancestors of the selection nodes can always be preserved when applying the conditioning operation.

We can see the rung-one and rung-two semantics of an SCM $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ as a collection of distributions $\mathcal{C}^V = \left\{\mathrm{P}^{[x_T]}(X_{V\setminus T}) : T \subseteq V, x_T \in \mathcal{X}_T\right\}$, which satisfies some constraints given by the SCM $M$, i.e., $\mathrm{P}^{[x_T]}(X_{V\setminus T}) = \mathrm{P}_M(X_{V\setminus T} \mid \mathrm{do}(X_T = x_T))$ for all $T \subseteq V$ and $x_T \in \mathcal{X}_T$.

Let

$$\mathcal{C}^{V\setminus S} \coloneqq \left(\mathrm{P}^{[x_T]}(X_{V\setminus(S\cup T)}) : T \subseteq V \setminus S, x_T \in \mathcal{X}_T, \mathrm{P}_M(X_S \in \mathcal{S} \mid \mathrm{do}(X_T = x_T)) > 0\right)$$

where

$$\mathrm{P}^{[x_T]}(X_{V\setminus(S\cup T)}) \coloneqq \frac{\mathrm{P}_M(X_{V\setminus(T\cup S)}, X_S \in \mathcal{S} \mid \mathrm{do}(X_T = x_T))}{\mathrm{P}_M(X_S \in \mathcal{S} \mid \mathrm{do}(X_T = x_T))}$$

for some simple s-SCM $(M, X_S \in \mathcal{S})$ with $M = (V, W, \mathcal{X}, \mathrm{P}, f)$. Note that $\mathcal{C}^{V\setminus S}$ is just the collection of observational and interventional distributions induced by the simple s-SCM $(M, X_S \in \mathcal{S})$. Now Question Q1 can be rephrased as: given $\mathcal{C}^{V\setminus S}$ defined above, can we *always* find a simple SCM $\widetilde{M}$ with endogenous nodes $O \coloneqq V \setminus S$ such that

$$\forall T \subseteq O, x_T \in \mathcal{X}_T \quad \mathrm{P}^{[x_T]}(X_{O\setminus T}) = \mathrm{P}_{\widetilde{M}}(X_O \mid \mathrm{do}(X_T = x_T))? \tag{6}$$

The answer is *No*:[32]

**Proposition D.1.** *There exists a simple s-SCM $(M, X_S \in \mathcal{S})$ such that it is impossible to find a simple SCM $\widetilde{M}$ with endogenous nodes $O \coloneqq V \setminus S$ such that Equation (6) holds.*

*Proof.* We first show that in general there is no acyclic SCM $\widetilde{M}$ such that equation (6) holds and second show that in general there is no simple (even cyclic) SCM satisfying equation (6). This can be summarized by Figure 23.

[32] This is also related to Lauritzen (1998), who shows that any hierarchical model can be generated from graphical models represented by a DAG with selection.

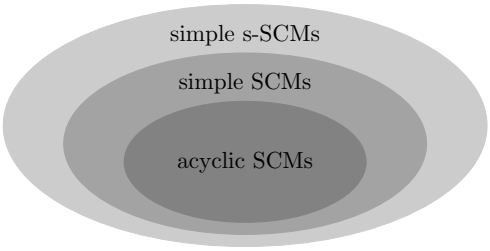
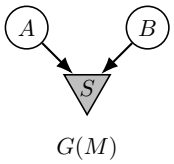



Figure 23: Venn diagram for different causal modeling classes with rung-two information.

**No acyclic SCM $\widetilde{M}$ satisfying equation (6).** Let $M$ be an acyclic SCM with $X_S = X_A + X_B$, $X_A = E_A \sim \mathrm{Uni}[0,1]$, $X_B = E_B \sim \mathrm{Uni}[0,1]$, and selection mechanism $X_S \geq 0.8$, whose graph is shown in Figure 24.




$G(M)$

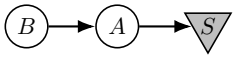
Figure 24: The causal graphs of the SCM $M$ with selection $X_S \geq 0.8$.

We shall explain the non-existence of an acyclic SCM $\widetilde{M}$ satisfying equation (6) by contradiction. Assume that there exists an acyclic SCM $\widetilde{M}$ satisfying equation (6). For $x_A \in [0,1]$ and $x_B \in [0,1]$, we require $\mathrm{P}_{\widetilde{M}}(X_B \mid \mathrm{do}(X_A = x_A)) = \mathrm{P}_M(X_B \mid \mathrm{do}(X_A = x_A), X_S \geq 0.8)$ and $\mathrm{P}_{\widetilde{M}}(X_A \mid \mathrm{do}(X_B = x_B)) = \mathrm{P}_M(X_A \mid \mathrm{do}(X_B = x_B), X_S \geq 0.8)$. Since $\mathrm{P}_M(X_B \mid \mathrm{do}(X_A = x_A), X_S \geq 0.8)$ and $\mathrm{P}_M(X_A \mid \mathrm{do}(X_B = x_B), X_S \geq 0.8)$ are not constant in $x_A$ and $x_B$ respectively, we must have a directed edge from $A$ to $B$ and a directed edge from $B$ to $A$ in the causal graph $G(\widetilde{M})$. Hence, $\widetilde{M}$ cannot be an acyclic SCM.

**No simple (even cyclic) SCM $\widetilde{M}$ satisfying equation (6).** Let $M$ be an SCM that satisfies

$$\begin{aligned} \mathrm{P}_M(X_B) &= 0.9\delta_0 + 0.1\delta_1, \\ \mathrm{P}_M(X_A \mid \mathrm{do}(X_B = 0)) &= 0.6\delta_0 + 0.4\delta_1, \ \mathrm{P}_M(X_A \mid \mathrm{do}(X_B = 1)) = 0.1\delta_0 + 0.9\delta_1, \\ \mathrm{P}_M(X_S \mid \mathrm{do}(X_A = 0)) &= 0.9\delta_0 + 0.1\delta_1, \ \mathrm{P}_M(X_S \mid \mathrm{do}(X_A = 1)) = 0.1\delta_0 + 0.9\delta_1, \end{aligned}$$

selection mechanism $X_S = 1$, and graph $G(M)$ shown in Figure 25.




$G(M)$

Figure 25: The causal graph of the SCM $M$ with selection $X_S = 1$.

Manski and Nagin (1998) proved a "natural" bound and we point out here that it also holds for simple SCMs. More specifically, for an arbitrary simple SCM $\widetilde{M}$ with discrete endogenous

variables $\{X_A, X_B\}$, it must hold that $\mathrm{P}_{\widetilde{M}}(X_A = x_A, X_B) \leq \mathrm{P}_{\widetilde{M}}(X_B \mid \mathrm{do}(X_A = x_A))$ for all $x_A \in \mathcal{X}_A$.[33] Indeed, by the consistency $X_B = X_B(X_A)$ a.s. (see also Forré and Mooij (2025, Proposition 7.5.1)) and elementary probability theory, we have

$$\begin{aligned}\mathrm{P}_{\widetilde{M}}(X_A = x_A, X_B) &= \mathrm{P}_{\widetilde{M}}(X_A = x_A, X_B(x_A)) \\ &\leq \mathrm{P}_{\widetilde{M}}(X_B(x_A)) = \mathrm{P}_{\widetilde{M}}(X_B \mid \mathrm{do}(X_A = x_A)).\end{aligned}$$

Assume that $\widetilde{M}$ is a simple SCM satisfying equation (6). Then by requiring

$$\mathrm{P}_{\widetilde{M}}(X_A = 1, X_B = 1) = \mathrm{P}_M(X_A = 1, X_B = 1 \mid X_S = 1)$$

and

$$\mathrm{P}_{\widetilde{M}}(X_B = 1 \mid \mathrm{do}(X_A = 1)) = \mathrm{P}_M(X_B = 1 \mid \mathrm{do}(X_A = 1), X_S = 1),$$

we have

$$\mathrm{P}_{\widetilde{M}}(X_A = 1, X_B = 1) \approx 0.176 > 0.1 = \mathrm{P}_{\widetilde{M}}(X_B = 1 \mid \mathrm{do}(X_A = 1)),$$

which contradicts the natural bound $\mathrm{P}_{\widetilde{M}}(X_A = 1, X_B = 1) \leq \mathrm{P}_{\widetilde{M}}(X_B = 1 \mid \mathrm{do}(X_A = 1))$ that the simple SCM $\widetilde{M}$ must satisfy. □

**Remark D.2** (Interventions on ancestors of selection nodes)**.** Let $T \subseteq \mathrm{Anc}_{G(M)}(S)$. The above theorem tells us that in general $\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T)) \neq \mathrm{P}_M(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S})$, since $\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T)) = \mathrm{P}_M(X_{O\setminus T}(x_T) \mid X_S \in \mathcal{S})$ and $\mathrm{P}_M(X_{O\setminus T}(x_T) \mid X_S \in \mathcal{S}) \neq \mathrm{P}_M(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S})$ in general. However, in some cases, we can infer that $\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T)) = \mathrm{P}_M(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S})$. For example, if $\mathrm{P}_M(X_S \in \mathcal{S} \mid \mathrm{do}(X_T = x_T)) = \mathrm{P}_M(X_S \in \mathcal{S})$ then we can conclude that $\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T)) = \mathrm{P}_M(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S})$. As another example, if we know that the second rule or the third rule of do-calculus applies to $\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), X_C)$ and $\mathrm{P}_M(X_{O\setminus T}, X_S \in \mathcal{S} \mid \mathrm{do}(X_T = x_T), X_C)$ to reduce $\mathrm{do}(X_T)$ to given $X_T$ or eliminate $\mathrm{do}(X_T)$ entirely, then we have the equality $\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T)) = \mathrm{P}_M(X_{O\setminus T} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S})$ under the assumption of discreteness and positivity. Consider a concrete toy example where we have an SCM with causal graph $S \leftarrow T \rightarrow O$. Under discreteness and positivity assumption, we have

$$\begin{aligned}\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_O \mid \mathrm{do}(X_T = x_T)) &= \mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_O \mid X_T = x_T) \\ &= \mathrm{P}_M(X_O \mid X_T = x_T, X_S \in \mathcal{S}) \\ &= \frac{\mathrm{P}_M(X_O, X_S \in \mathcal{S} \mid X_T = x_T)}{\mathrm{P}_M(X_S \in \mathcal{S} \mid X_T = x_T)} \\ &= \mathrm{P}_M(X_O \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S}).\end{aligned}$$

It might be possible to find some interesting conditions to guarantee this, and therefore in the given setting one does not need to treat the ancestors of selection nodes differently than the non-ancestors.

[33] This inequality is interpreted as $\mathrm{P}_{\widetilde{M}}(X_A = x_A, X_B = x_B) \leq \mathrm{P}_{\widetilde{M}}(X_B = x_B \mid \mathrm{do}(X_A = x_A))$ for all $x_B \in \mathcal{X}_B$.

## D.2 Other variants of conditioning operation for SCMs

Definition 3.5 is not the only possible way to define a "conditioned SCM". In this section, we explore some other possibilities of conditioning operations such as different decompositions of exogenous nodes, and conditioning for Causal Bayesian Networks.[34]

### D.2.1 Different decomposition of exogenous nodes

Why do we care about decomposition of exogenous variables and make sure that the new coarse-grained variables are mutually independent given selection? That is because we want to have a Markov property of the causal graphs of our SCMs (so that we can apply do-calculus, and so on) and without mutual independence of the exogenous variables this may fail.

For example, in some literature (such as Bareinboim et al. (2022)), SCMs are not required to have mutually independent exogenous random variables but can have any exogenous probability distribution $\mathrm{P}(X_W)$ on $X_W$. A bidirected edge is drawn between two endogenous variables $X_{v_1}$ and $X_{v_2}$ if they share the same exogenous variables or their exogenous parents are correlated according to $\mathrm{P}(X_W)$. It seems that if we adopt this framework, then we just need to update the exogenous distribution and not to merge exogenous variables when defining a conditioning operation. Consider the "SCM"

$$M: \begin{cases} \mathrm{P}(E_1, E_2, E_3) = \frac{1}{4}\delta_{000} + \frac{1}{4}\delta_{011} + \frac{1}{4}\delta_{101} + \frac{1}{4}\delta_{110} \\ X_1 = E_1 \\ X_2 = E_2 \\ X_3 = E_3. \end{cases}$$

According to the above definition, this "SCM" $M$ would have graph $G(M)$ shown in Figure 26.

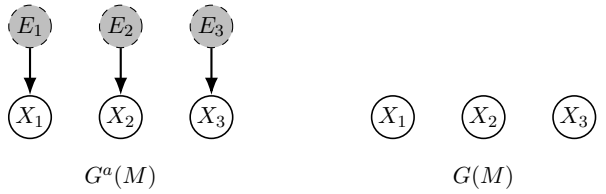



Figure 26: The "causal" graph of the "SCM" $M$.

The graph is of this form because $E_i \perp\!\!\!\perp E_j$ for $i, j = 1, 2, 3$ and $i \neq j$. However, although the graph implies that $X_1, X_2 \overset{d}{\underset{G(M)}{\perp}} X_3$, we know from the model $M$ that

$$X_1, X_2 \underset{\mathrm{P}_M(X_1,X_2,X_3)}{\not\perp\!\!\!\perp} X_3.$$

Therefore, the Markov property does not hold, and all the results based on it may not hold either, such as the back-door criterion and Pearl's do-calculus.

It is worth mentioning that one can have different decompositions of the exogenous nodes of the conditioned SCMs. One extreme is to merge all the exogenous nodes into one single

[34] Conditioning on a variable that we want to observe is discussed in Section 3.2.2.

node, which results in the "coarsest" model without much information. The other extreme is to consider the "finest" decomposition of the exogenous nodes under the current framework of SCMs, which results in "most fine-grained" models with as much information as possible.

In Definition 3.5, we used a "finest" decomposition. We can consider any decomposition of label set $W$ given $X_S \in \mathcal{S}$ that is coarser than the one given in Definition 3.5. There are two such examples that appear natural, but one should note that the properties of these two operations are slightly different from the ones of Definition 3.5.

**Definition D.3** (Conditioned SCMs II)**.** *In the setting of Definition 3.5, we define* $M_{|X_S \in \mathcal{S}} = (\widehat{V}, \bar{W}, \widehat{\mathcal{X}}, \widehat{f}, \widehat{\mathrm{P}})$ *where* $\widehat{V}, \widehat{\mathcal{X}}, \widehat{f}, \widehat{\mathrm{P}}$ *are the same as Definition 3.5, while*

$$\bar{W} := \{W \cap \mathrm{Anc}_{G^a(M_{\setminus(V \setminus S)})}(S)\} \dot{\cup} (W \setminus \mathrm{Anc}_{G^a(M_{\setminus(V \setminus S)})}(S)).$$

**Definition D.4** (Conditioned SCMs III)**.** *In the setting of Definition 3.5, we define* $M_{|X_S \in \mathcal{S}} = (\widehat{V}, \bar{\bar{W}}, \widehat{\mathcal{X}}, \widehat{f}, \widehat{\mathrm{P}})$ *where* $\widehat{V}, \widehat{\mathcal{X}}, \widehat{f}, \widehat{\mathrm{P}}$ *are the same as Definition 3.5, while*

$$\bar{\bar{W}} := \{W \cap \mathrm{Anc}_{G^a(M)}(S)\} \dot{\cup} (W \setminus \mathrm{Anc}_{G^a(M)}(S)).$$

For Definition D.4, we do not have Proposition 3.18, but we can obtain the commutativity of marginalization and conditioning up to counterfactual equivalence (see an old version of the current article (Chen et al., 2024b)).

### D.2.2 conditioning operation for causal Bayesian networks

Given a Causal Bayesian Network $N = (G = (V, E^d), \{\mathrm{P}(X_v \mid \mathrm{do}(X_{\mathrm{Pa}_G(v)}))\}_{v \in V})$, we can use the deterministic representation of Markov kernels to construct an SCM $M_N = (V, W, \mathcal{X}, \mathrm{P}, f)$ (i.e., there exist a uniformly distributed random variable $U_v$ and a measurable function $f_v$ such that $\mathrm{P}(X_v \mid \mathrm{do}(X_{\mathrm{Pa}_G(v)})) = f_v(X_{\mathrm{Pa}_G(v)}, U_v)_* \mathrm{P}(U_v)$, see, e.g., Bogachev (2007, Proposition 10.7.6)). We can then apply the conditioning operation for $M_N$ and then transform the conditioned SCM back to get a conditioned Causal Bayesian Network with latent variables.

## D.3 Conditioning operation for SCMs with inputs

In this section, we extend the definition of the conditioning operation (Definition 3.5) on SCMs to SCMs with input nodes, which we call **iSCMs** and are introduced in Forré and Mooij (2025). The difference between iSCMs and SCMs is that iSCMs have exogenous (non-stochastic) input variables in addition to endogenous and exogenous random variables. Note that such an extension of conditioning operation is not straightforward due to interactions between non-stochastic variables and stochastic variables (cf. Remark D.6 and Definition D.9).

**Definition D.5** (SCMs with input nodes (iSCMs))**.** *A* ***Structural Causal Model with input nodes (iSCM)*** *is a tuple* $M = (J, V, W, \mathcal{X}, \mathrm{P}, f)$ *where* $J$ *represents the label set for* ***exogenous input (non-stochastic) variables****. Other components have the same definitions as their counterparts in Definition 2.1, except for* $\mathcal{X} = \prod_{i \in J \dot{\cup} V \dot{\cup} W} \mathcal{X}_i$.

All the definitions in Section 2 can be extended to iSCMs with minor modifications (see (Forré and Mooij, 2025) for more details. For example, an iSCM could induce a Markov kernel $\mathrm{P}_M(X_V \mid \mathrm{do}(X_J))$ in general and not merely a probability distribution $\mathrm{P}_M(X_V)$. A

solution function $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ has also arguments from $x_J \in \mathcal{X}_J$ and all quantifiers "for $\mathrm{P}(X_W)$-a.a. $x_W \in \mathcal{X}_W$, and all $x_V \in \mathcal{X}_V$" used in the relevant definition (e.g., essentially unique solution function) are replaced by "for all $x_J \in \mathcal{X}_J$, $\mathrm{P}(X_W)$-a.a. $x_W \in \mathcal{X}_W$, and all $x_V \in \mathcal{X}_V$" (note that the ordering of these quantifiers is important). We define $M_{\mathrm{do}(T)}$ as an iSCM where we transform $T$ from endogenous nodes to exogenous input nodes for $T \subseteq V$. This is in contrast to the case in SCMs where we always need to assign a specific value to the intervened variables.

The input variables can model hard/soft interventions, (non-stochastic) parameters for models, context variables, and regime indicators (Dawid, 2002, 2021) or say policy variables (Spirtes et al., 2001), etc. Conceptually, iSCMs model data-generating processes where first the values of $X_J$ are assigned (e.g., by the experimenter) and then the process in Algorithm 3.3 is implemented (function $g$ also depends on $X_J$). Mathematically, input variables can help rigorously develop a general measure-theoretic causal calculus (Forré, 2021; Forré and Mooij, 2025).[35] One common feature of all these concepts is that they are modeled by variables without probability distributions on them and, therefore, need a distinct treatment from ordinary random variables.

One convenient foundational framework for dealing with such non-stochastic and ordinary stochastic variables universally is named **transitional probability theory** in (Forré, 2021). We will use relevant concepts directly and refer the reader to Forré (2021) for more details.

**Remark D.6** (Interaction between non-stochastic variables and stochastic variables)**.** Given $X_S \in \mathcal{S}$, the exogenous distribution $\mathrm{P}(X_W)$ becomes an exogenous Markov kernel

$$\mathrm{P}_M(X_W \mid X_S \in \mathcal{S}, \mathrm{do}(X_J)) = \frac{\mathrm{P}_M(X_W, X_S \in \mathcal{S} \mid \mathrm{do}(X_J))}{\mathrm{P}_M(X_S \in \mathcal{S} \mid \mathrm{do}(X_J))},$$

where the exogenous distribution has a dependency on $X_J$. Since $\mathrm{P}_M(X_S \in \mathcal{S} \mid \mathrm{do}(X_J = x_J)) = 0$ might be possible for some $x_J \in \mathcal{X}_J$, this might require merging some exogenous input nodes and restricting $\mathcal{X}_J$.

To eliminate the "entanglement" between $X_J$ and $X_W$ mentioned in Remark D.6, we need Bogachev (2007, Theorem Proposition 10.7.6). It states that we can represent a Markov kernel as a deterministic map of a random variable.

**Proposition/Definition D.7** (Deterministic representation of Markov kernels)**.** Let $\mathcal{Z}$ be any measurable space, $\mathcal{X}$ be a standard measurable space, and $\mathrm{P}(X \mid Z) : \mathcal{Z} \to \mathcal{P}(\mathcal{X})$ be a Markov kernel. Then there exists a measurable function $R : \mathcal{Z} \times \mathcal{U} \to \mathcal{X}$ such that

$$\mathrm{P}(X \mid Z) = \mathrm{P}(R(Z, U) \mid Z),$$

where $\mathcal{U}$ is a measurable space and $U$ a random variable taking values in $\mathcal{U}$. We call $(U, R)$ a **deterministic representation of the Markov kernel** $\mathrm{P}(X \mid Z)$.

**Remark D.8.** One can take $\mathrm{P}(U)$ to be $\mathrm{Uni}([0,1]^n)$ or $\mathcal{N}(0, I_n)$. After fixing $U$, in general, the measurable function $R$ is *not* unique, injective, or surjective.

[35] When the variables are not discrete, a naive approach would not work. See the discussions in Forré and Mooij (2020) and Forré (2021).

The conditioning operation for iSCMs can be seen as the composition of (i) merging all the input nodes that are ancestors of the selection nodes and exogenous random nodes that are ancestors of the selection nodes, respectively, restricting the values that input variables can take given $X_S \in \mathcal{S}$, and then representing the corresponding conditioned Markov kernel deterministically to eliminate the "dependence" between input variables and exogenous random variables, and (ii) marginalizing out the selection variables.

Let $M = (J, V, W, \mathcal{X}, \mathrm{P}, f)$ be a simple iSCM (Forré and Mooij, 2025, Definition 6.6.5). Let $S \subseteq V$ be a subset of endogenous nodes and $\mathcal{S} \subseteq \mathcal{X}_S$ a measurable subset of values $X_S$ may take where there exists $x_J \in \mathcal{X}_J$ such that $\mathrm{P}_M(X_S \in \mathcal{S} \mid \mathrm{do}(X_J = x_J)) > 0$. Let $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ be the (essentially unique) solution function of $M$. Note that we can see $g_S$ as a map from $\mathcal{X}_{J \cap \mathrm{Anc}_{G(M)}(S)} \times \mathcal{X}_{W \cap \mathrm{Anc}_{G^a(M)}(S)}$ to $\mathcal{X}_S$. Write $O \coloneqq V \setminus S$, $B_1 \coloneqq \mathrm{Anc}_{G(M)}(S) \cap J$, and $B_2 \coloneqq \mathrm{Anc}_{G^a(M)}(S) \cap W$.

**Definition D.9** (Conditioned iSCM)**.** *Under the above setting, we define a conditioned iSCM* $M_{|X_S \in \mathcal{S}} \coloneqq \left(\widehat{J}, \widehat{V}, \widehat{W}, \widehat{\mathcal{X}}, \widehat{\mathrm{P}}, \widehat{f}\right)$ *by*

*(1)* $\widehat{J} \coloneqq \{\star_J\} \dot\cup (J \setminus B_1)$ *where* $\star_J = B_1$*;*

*(2)* $\widehat{W} \coloneqq \{\star_W\} \dot\cup (W \setminus B_2)$ *where* $\star_W = B_2$*;*

*(3)* $\widehat{V} \coloneqq V \setminus S$*;*

*(4)* $\mathcal{X}_{\widehat{J}} \coloneqq \mathcal{X}_{\star_J} \times \mathcal{X}_{J \setminus B_1}$ *where*

$$\mathcal{X}_{\star_J} \coloneqq \{x_{B_1} \in \mathcal{X}_{B_1} \mid \mathrm{P}_M(X_S \in \mathcal{S} \mid \mathrm{do}(X_{B_1} = x_{B_1})) > 0\}$$

*and* $\mathcal{X}_{\widehat{W}} \coloneqq \mathcal{X}_{\star_W} \times \mathcal{X}_{W \setminus B_2}$ *with* $\mathcal{X}_{\star_W} = [0,1]$ *and* $\mathcal{X}_{\widehat{V}} \coloneqq \mathcal{X}_{V \setminus S}$*;*

*(5)* $\widehat{\mathrm{P}}(X_{\widehat{W}}) \coloneqq \mathrm{P}(X_{W \setminus B_2}) \otimes \mathrm{P}(X_{\star_W})$ *with* $\mathrm{P}(X_{\star_W}) = \mathrm{Uni}([0,1])$*;*

*(6) causal mechanism:*

$$\widehat{f}(x_{\widehat{J}}, x_{\widehat{V}}, x_{\widehat{W}}) \coloneqq$$
$$f_O(x_{J \setminus B_1}, x_{\star_J}, x_O, g^S(x_{J \setminus B_1}, x_{\star_J}, x_O, x_{W \setminus B_2}, R(x_{\star_J}, x_{\star_W})), x_{W \setminus B_2}, R(x_{\star_J}, x_{\star_W})),$$

*where* $(X_{\star_W}, R)$ *is a deterministic representation of the Markov kernel* $\mathrm{P}_M(X_{B_2} \mid \mathrm{do}(X_{B_1}), X_S \in \mathcal{S})$ *and* $g^S : \mathcal{X}_J \times \mathcal{X}_O \times \mathcal{X}_W \to \mathcal{X}_S$ *is the (essentially unique) solution function of* $M$ *w.r.t.* $S$*.*

**Remark D.10.** Here we simply merge all the input node ancestors of $S$ and exogenous node ancestors of $S$, respectively. One can also derive a finer merging scheme, similar to what we did for Definition 3.5. For the sake of space, we did not spell out all the details. The essential point that we want to show in this subsection is how to eliminate the dependency between input variables and exogenous random variables given $X_S \in \mathcal{S}$.

One can develop a theory for the conditioning operation for iSCMs and show the corresponding properties in parallel to what we did in Section 3. We now define an operation on iSCMs called **exogenous (quasi-)pullback of iSCMs**, which defines formally the first step of the conditioning operation for iSCMs. Because of the nice properties of marginalization

Forré and Mooij (2025, Section 6.8), to show properties of the conditioning operation for iSCMs it suffices to show that the corresponding properties of exogenous quasi-pullback of iSCMs hold. See Forré and Mooij (2025, Section 8.3.2) for some properties of exogenous pullback of iSCMs. Note that merging exogenous nodes is a special case of exogenous (quasi-)pullback of iSCMs.

**Definition D.11** (Exogenous quasi-pullback iSCMs)**.** *Let $M = (J, V, W, \mathcal{X}, \mathrm{P}, f)$ be an iSCM. Let $\widetilde{M} = (\widetilde{J}, V, \widetilde{W}, \widetilde{\mathcal{X}}, \widetilde{\mathrm{P}}, \widetilde{f})$ be an iSCM. Let $\Phi_J : \mathcal{X}_{\widetilde{J}\cup\widetilde{W}} \to \mathcal{X}_J$ be a measurable mapping that does not depend on the $X_{\widetilde{W}}$-component, and let $\Phi_W : \mathcal{X}_{\widetilde{J}\cup\widetilde{W}} \to \mathcal{X}_W$ be a measurable mapping such that $\mathrm{Q}(X_W \mid X_{\widetilde{J}}) := (\Phi_W)_* \left(\delta(X_{\widetilde{J}} \mid X_{\widetilde{J}}) \otimes \widetilde{\mathrm{P}}(X_{\widetilde{W}})\right) \ll \mathrm{P}(X_W)$, i.e., for all $x_{\widetilde{J}} \in \mathcal{X}_{\widetilde{J}}$, and for every measurable subset $\mathcal{A} \subseteq \mathcal{X}_W$, $\mathrm{P}(X_W \in \mathcal{A}) = 0$ implies that $\mathrm{Q}(X_W \in \mathcal{A} \mid X_{\widetilde{J}} = x_{\widetilde{J}}) = 0$. Assume that*

$$\widetilde{f}(x_{\widetilde{J}}, x_V, x_{\widetilde{W}}) = f\left(\Phi_J(x_{\widetilde{J}}), x_V, \Phi_W(x_{\widetilde{J}}, x_{\widetilde{W}})\right).$$

*Then we call $\Phi = (\Phi_J, \Phi_W) : \mathcal{X}_{\widetilde{J}\cup\widetilde{W}} \to \mathcal{X}_J \times \mathcal{X}_W$ an **exogenous quasi-pullback function of** $M$ and $\widetilde{M}$ an **exogenous quasi-pullback iSCM of** $M$ **associated with** $(\Phi, \widetilde{\mathrm{P}})$. We denote $\widetilde{M}$ by $M_{\mathrm{ep}(\Phi,\widetilde{\mathrm{P}})}$.*

It is easy to see that $M_{|X_S \in \mathcal{S}} = \left(M_{\mathrm{ep}(\Phi,\widehat{\mathrm{P}})}\right)_{\setminus S}$ is a marginalized exogenous quasi-pullback iSCM of $M$ associated with the exogenous quasi-pullback function $\Phi = (\Phi_J, \Phi_W)$ and distribution $\widehat{\mathrm{P}}(X_{\widehat{W}})$ where

$$\begin{aligned} &\Phi_J : \mathcal{X}_{\widehat{J}\cup\widehat{W}} \to \mathcal{X}_J, \Phi_J(x_{\widehat{J}}, x_{\widehat{W}}) := x_J \\ \text{and } &\Phi_W : \mathcal{X}_{\widehat{J}\cup\widehat{W}} \to \mathcal{X}_W, \Phi_W(x_{\widehat{J}}, x_{\widehat{W}}) := (x_{W\setminus B_2}, R(x_{\star_J}, x_{\star_W})). \end{aligned}$$

The graphical conditioning operation for DMGs with input nodes can be defined similarly to Definition 3.23 by (i) merging all the input node ancestors of $S$, (ii) adding bidirected edges to output nodes that are ancestors or siblings of ancestors of $S$, and (iii) marginalizing out $S$. In the graph, we make both the merged input node (that corresponds with those input nodes that were ancestors of $S$) and the output nodes that were ancestors of $S$ dashed. One can develop a theory for this operation and show the corresponding properties in parallel to what we did in Section 3.3. Note that one needs to replace stochastic conditional independence and the usual graphical $\sigma$-separation with transitional conditional independence (Forré, 2021, Definition 3.1) and a nuanced graphical separation (Forré, 2021, Definition 5.9), respectively.

We leave further exploration of the properties of the conditioning operations for iSCMs and on DMGs with input nodes for future work.

## References


Abouei, A. M., Mokhtarian, E., and Kiyavash, N. (2024a). s-id: Causal effect identification in a sub-population. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):20302–20310. ↑3, 32, 33

Abouei, A. M., Mokhtarian, E., Kiyavash, N., and Grossglauser, M. (2024b). Causal effect identification in a sub-population with latent variables. *arXiv.org preprint, arXiv:2405.14547 [cs.LG]*. ↑32

Badhane, S., Mooij, J. M., Boeken, P., and Zoeter, O. (2025). Revisiting the Berkeley admissions data: Statistical tests for causal hypotheses. *arXiv.org preprint*, arXiv:2502.10161 [stat.ME]. ↑34, 35

Balke, A. and Pearl, J. (1994). Counterfactual probabilities: computational methods, bounds and applications. In *Proceedings of the 10th International Conference on Uncertainty in Artificial Intelligence*, UAI'94, page 46–54, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. ↑34

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176. ↑34

Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. (2022). *On Pearl's Hierarchy and the Foundations of Causal Inference*, page 507–556. Association for Computing Machinery, New York, NY, USA, 1 edition. ↑59

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. ↑3, 8

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. ↑8

Beckers, S. and Halpern, J. Y. (2019). Abstracting causal models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2678–2685. ↑9, 10

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

Blom, T., Bongers, S., and Mooij, J. M. (2020). Beyond structural causal models: Causal constraints models. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference (UAI-19)*, volume 115 of *Proceedings of Machine Learning Research*, pages 585–594. PMLR. ↑4, 20

Bogachev, V. I. (2007). *Measure Theory*, volume 1. Springer. ↑60, 61

Bongers, S., Blom, T., and Mooij, J. M. (2022). Causal modeling of dynamical systems. *arXiv.org preprint*, arXiv:1803.08784v4 [cs.AI]. ↑9

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. ↑3, 10, 11, 12, 19, 21, 39, 40, 41, 42, 46, 49, 52, 53

Chen, L., Fritz, T., Gonda, T., Klingler, A., and Lorenzin, A. (2024a). The Aldous–Hoover theorem in categorical probability. *arXiv.org preprint, arXiv:2411.12840 [math.ST]*. ↑38

Chen, L., Zoeter, O., and Mooij, J. M. (2024b). Modeling latent selection with structural causal models. *arXiv.org preprint, arXiv:2401.06925v2 [cs.AI]*. ↑60

Chiappa, S. (2019). Path-specific counterfactual fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7801–7808. ↑35

Cooper, G. F. (1995). Causal discovery from data in the presence of selection bias. In Fisher, D. and Lenz, H.-J., editors, *Pre-proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, volume R0 of *Proceedings of Machine Learning Research*, pages 140–150. PMLR. Reissued by PMLR on 01 May 2022. ↑3, 7, 9

Correa, J. and Bareinboim, E. (2017). Causal effect identification by adjustment under confounding and selection biases. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1). ↑31

Correa, J. and Bareinboim, E. (2020). A calculus for stochastic interventions:causal effect identification and surrogate experiments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(06):10093–10100. ↑11

Correa, J., Lee, S., and Bareinboim, E. (2021). Nested counterfactual identification from arbitrary surrogate experiments. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6856–6867. Curran Associates, Inc. ↑43

Daniel, R. M., Kenward, M. G., Cousens, S. N., and De Stavola, B. L. (2012). Using causal diagrams to guide analysis in missing data problems. *Statistical methods in medical research*, 21(3):243–256. ↑3

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189. ↑10, 28, 61

Dawid, A. P. (2021). Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9(1):39–77. ↑10, 28, 61

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

Forré, P. (2021). Transitional conditional independence. *arXiv.org preprint, arXiv:2104.11547 [math.ST]*. ↑28, 61, 63

Forré, P. and Mooij, J. M. (2017). Markov properties for graphical models with cycles and latent variables. *arXiv.org preprint*, arXiv:1710.08775 [math.ST]. ↑3, 23, 40, 41

Forré, P. and Mooij, J. M. (2018). Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pages 269–278. ↑8, 26

Forré, P. and Mooij, J. M. (2020). Causal calculus in the presence of cycles, latent confounders and selection bias. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, pages 71–80. ↑28, 61

Forré, P. and Mooij, J. M. (2025). A mathematical introduction to causality. ↑30, 48, 51, 58, 60, 61, 62, 63

Fritz, T. (2020). A synthetic approach to markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239. ↑38

Fritz, T., Gonda, T., Lorenzin, A., Perrone, P., and Mohammed, A. S. (2025). Empirical measures and strong laws of large numbers in categorical probability. *arXiv.org preprint, arXiv:2503.21576 [math.PR]*. ↑38

Fritz, T., Gonda, T., and Perrone, P. (2021). De finetti's theorem in categorical probability. *Journal of Stochastic Analysis*, 2(4). ↑38

Fritz, T. and Klingler, A. (2023). The d-separation criterion in categorical probability. *Journal of Machine Learning Research*, 24(46):1–49. ↑38

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

Geiger, A., Potts, C., and Icard, T. F. (2023). Causal abstraction for faithful model interpretation. *arXiv.org preprint*, arXiv:2301.04709 [cs.AI]. ↑9

Geiger, D., Verma, T., and Pearl, J. (1990). Identifying independence in bayesian networks. *Networks*, 20(5):507–534. ↑24

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161. ↑3

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC. ↑9, 12

Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615–625. ↑3, 9

Huang, Y. and Valtorta, M. (2006). Pearl's calculus of intervention is complete. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, page 217–224. ↑32

Huang, Y. and Valtorta, M. G. (2008). On the completeness of an identifiability algorithm for semi-markovian models. *Annals of Mathematics and Artificial Intelligence*, 54:363–408. ↑3, 32

Hyttinen, A., Eberhardt, F., and Järvisalo, M. (2014). Constraint-based causal discovery: Conflict resolution with answer set programming. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, page 340–349. ↑8

Imbens, G., Angrist, J., and Graddy, K. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Review of Economic Studies*, 67, July:499–527. ↑34

Jaynes, E. T. (2003). *Paradoxes of probability theory*, page 451–489. Cambridge University Press. ↑42

Kivva, Y., Etesami, J., and Kiyavash, N. (2023). On identifiability of conditional causal effects. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*. ↑32

Kusner, M., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4069–4079, Red Hook, NY, USA. Curran Associates Inc. ↑9, 35, 43

Lauritzen, S. (1998). *Generating mixed hierarchical interaction models by selection*. Number R-98-2009 in Research Report Series. Aalborg Universitetsforlag, Denmark. ↑56

Lorenz, R. and Tull, S. (2023). Causal models in string diagrams. *arXiv.org preprint, arXiv:2304.07638 [cs.LO]*. ↑38

Lu, H., Cole, S. R., Howe, C. J., and Westreich, D. (2022). Toward a clearer definition of selection bias when estimating causal effects. *Epidemiology*, 33(5):699–706. ↑3, 43

Maathuis, M. H. and Colombo, D. (2015). A generalized back-door criterion. *The Annals of Statistics*, 43(3):1060 – 1088. ↑32

Manski, C. F. and Nagin, D. S. (1998). Bounding disagreements about treatment effects: A case study of sentencing and recidivism. *Sociological Methodology*, 28(1):99–137. ↑57

Mathur, M. B. and Shpitser, I. (2024). Simple graphical rules for assessing selection bias in general-population and selected-sample treatment effects. *American Journal of Epidemiology*, 194(1):267–277. ↑19

Mohammed, A. S. (2025). Partializations of Markov categories. *arXiv.org preprint, arXiv:2509.05094 [math.CT]*. ↑38

Mooij, J. M. and Claassen, T. (2020). Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. In Peters, J. and Sontag, D., editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI-20)*, volume 124, pages 1159–1168. PMLR. ↑8, 26

Munafo, M. R., Tilling, K., Taylor, A. E., Evans, D. M., and Smith, G. D. (2016). Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology*, 47:226 – 235. ↑3

Nabi, R. and Shpitser, I. (2018). Fair inference on outcomes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). ↑35

Pearl, J. (1995a). Causal diagrams for empirical research. *Biometrika*, 82(4):669–710. With discussion. ↑32

Pearl, J. (1995b). On the testability of causal models with latent and instrumental variables. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, page 435–443. ↑34

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 411–420. PMLR. ↑34

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition. ↑3, 6, 9, 26, 30, 31, 34, 37, 43

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459—481. ↑34, 35

Pearl, J. (2015). Conditioning on post-treatment variables. *Journal of Causal Inference*, 3(1):131–137. ↑10, 29

Pearl, J. (2019). On the interpretation of do(x). *Journal of Causal Inference*, 7(1). ↑10

Peters, J., Bauer, S., and Pfister, N. (2022). Causal models for dynamical systems. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 671–690. ↑9

Reichenbach, H. (1956). *The Direction of Time*. Dover Publications, Mineola, N.Y. ↑30

Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157. ↑6, 8, 23

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. (2023). Nested Markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1):334–361. ↑3, 33

Richardson, T. S. and Robins, J. M. (2013a). Single world intervention graphs: a primer. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=07bbcb458109d2663acc0d098e8913892389a2a7. ↑8, 9, 37

Richardson, T. S. and Robins, J. M. (2013b). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. ↑43

Richardson, T. S. and Spirtes, P. (2002). Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030. ↑4, 8, 33, 44

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3:143–155. ↑34

Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. (2017). Causal consistency of structural equation models. In *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI-17)*. ↑9, 10

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701. ↑12

Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science*, 37(6):1011–1035. ↑34, 35

Shpitser, I. and Pearl, J. (2006a). Identification of conditional interventional distributions. In *Proceedings of the 22d Conference on Uncertainty in Artificial Intelligence*, page 437–444. ↑32

Shpitser, I. and Pearl, J. (2006b). Identification of joint interventional distributions in recursive semi-markovian causal models. Proceedings of the National Conference on Artificial Intelligence, pages 1219–1226. ↑32

Smith, L. H. (2020). Selection mechanisms and their consequences: understanding and addressing selection bias. *Current Epidemiology Reports*, 7:179–189. ↑3, 35, 43

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. MIT Press. ↑24, 26, 61

Spirtes, P., Meek, C., and Richardson, T. (1995). Causal inference in the presence of latent variables and selection bias. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, page 499–506. ↑8, 26, 33

Spirtes, P., Meek, C., and Richardson, T. (1999). An algorithm for causal inference in the presence of latent variables and selection bias. In *Computation, Causation, and Discovery*. AAAI Press. ↑26

Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, page 567–573, USA. American Association for Artificial Intelligence. ↑3, 32

Van Himbeeck, T., Bohr Brask, J., Pironio, S., Ramanathan, R., Sainz, A. B., and Wolfe, E. (2019). Quantum violations in the Instrumental scenario and their relations to the Bell scenario. *Quantum*, 3:186. ↑34

VanderWeele, T. J. and Shpitser, I. (2013). On the definition of a confounder. *The Annals of Statistics*, 41(1):196–220. ↑13

Versteeg, P., Mooij, J., and Zhang, C. (2022). Local constraint-based causal discovery under selection bias. In Schölkopf, B., Uhler, C., and Zhang, K., editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 840–860. PMLR. ↑9

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. ↑36

Wald, A. (1943). A method of estimating plane vulnerability based on damage of survivors. Technical report, Statistical Research Group, Columbia University. Reprinted as Center for Naval Analyses Research Contribution CRC-432, 1980. ↑3

Wang, Y. S. and Drton, M. (2023). Causal discovery with unobserved confounding and non-gaussian data. *Journal of Machine Learning Research*, 24(271):1–61. ↑33

Williams, D. (1991). *Probability with Martingales*. Cambridge University Press. ↑54

Wolfe, E., Spekkens, R. W., and Fritz, T. (2019). The inflation technique for causal inference with latent variables. *Journal of Causal Inference*, 7(2):20170020. ↑33

Zhang, J. (2008). Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474. ↑8

Zhang, J. and Bareinboim, E. (2018). Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32. ↑9, 35, 43

Zhang, J. and Bareinboim, E. (2021). Bounding causal effects on continuous outcome. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):12207–12215. ↑34

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. ↑3