A Mathematical Introduction to Causality

Lecture Notes

Patrick Forré & Joris M. Mooij

June 13, 2025

Foreword

Causality is a broad topic, and these lecture notes cover only part of it. They originated over a period of four years as a by-product of the course on causality we taught for MSc. mathematics students. Our aim was to give a mathematically rigorous exposition of the graphical account to causal modeling, reasoning and inference, in the spirit of Wright, Spirtes, Glymour, Scheines, Pearl, and many others. Since there seemed to be no book or lecture notes out there that would fit our purpose, we decided to write our own.

The amount of material has grown over the years, and is still growing. We treat causal modeling with causal Bayesian networks (also known as 'DAGs') and structural causal models. Some unique features of our exposition are:

1. we have extended the standard formalisms with *input nodes* to enable a measure-theoretically rigorous treatment of the families of probability distributions that result from perfect interventions;

2. we allow for (sufficiently weak) cycles in structural causal models;

3. we have taken lots of care to provide a high level of mathematical rigor and consistency;

4. we emphasize the central role played by the Markov property in the theory.

Our treatment is self-contained: we start with the basic definitions (with as prerequisites only basic measure theory and probability theory), and derive everything that is necessary to prove the validity of Markov properties, the do-calculus, adjustment criteria, all the way up to extended versions of the ID algorithm and the FCI algorithm. We show how—with relatively little extra work—the framework of causal modeling with directed acyclic graphs can be extended to directed graphs that may have cycles.

While the advantages of mathematical rigor should be obvious, the price paid is that the non-trivial conceptual issues are sometimes clouded by technicalities. We believe that our treatment fills a much needed gap in the literature on causality, and consider it complementary to the many existing writings on similar topics (which often focus more on concepts and less on mathematical rigor).

We are indebted to our teaching assistants Leon Lang, Philip Boeken, Pim de Haan and Noud de Kroon for providing feedback and for spotting several errors in earlier drafts. While the current version undoubtedly still contains mistakes, we believe that it is now ready for wider exposure. We appreciate any feedback that the reader may have, be it on content, typos, or (we hope not) more serious mistakes.

Joris Mooij & Patrick Forré
Amsterdam
June 2025

Table of Contents

1. Experimental Causal Discovery

1.1. Types of Correlations

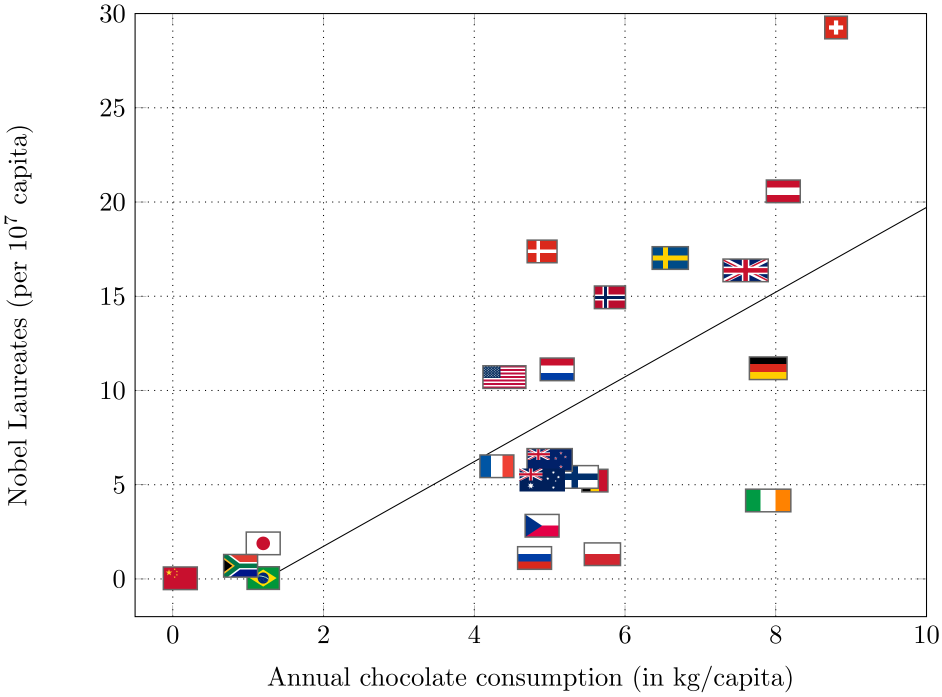


Figure 1: Statisically significant correlation between chocolate consumption and Nobel prizes (correlation coefficient $R = 0.69$, p-value for zero correlation $p = 0.0004$.[1]

Explanation 1.1.1. *What conclusions can we draw from this? Where does the correlation come from? Would the correlation hold under different conditions/circumstances? There are several explanations/stories that one could build around the observed correlation between the number of Nobel prizes N and the chocolate consumption per capita C:*

a) N causes C: "Nobel prize winning countries like to celebrate with chocolate consumption."

b) N is an effect of C: "Chocolate contains brain enhancing chemicals."

c) Feedback between N and C: Both stories hold.

[1]This figure is inspired by [Mes12, Figure 1]. We made a similar visualization, but using newer data from https://www.theobroma-cacao.de/wissen/wirtschaft/international/konsum on chocolate consumption in 2017, and from https://en.wikipedia.org/wiki/List_of_countries_by_Nobel_laureates_per_capita on scientific Nobel laureates until 2019.

d) Selection bias between N and C: "N and C are actually independent, but the data used was biased. For example, only countries that end up close to the diagonal line in Figure 1 were included in the plot."

e) Functional constraints between N and C: "International regulations make sure that Nobel prizes and chocolate imports are subtracted/added if they violate a linear relationship."

f) N and C have a common cause: "The wealth of a country determines both, how much money goes to science and also how much people can spend on chocolate."

g) Other explanations, e.g. measurement error, statistical coincidence, other forms of spurious correlations, combinations of all of these, etc.?

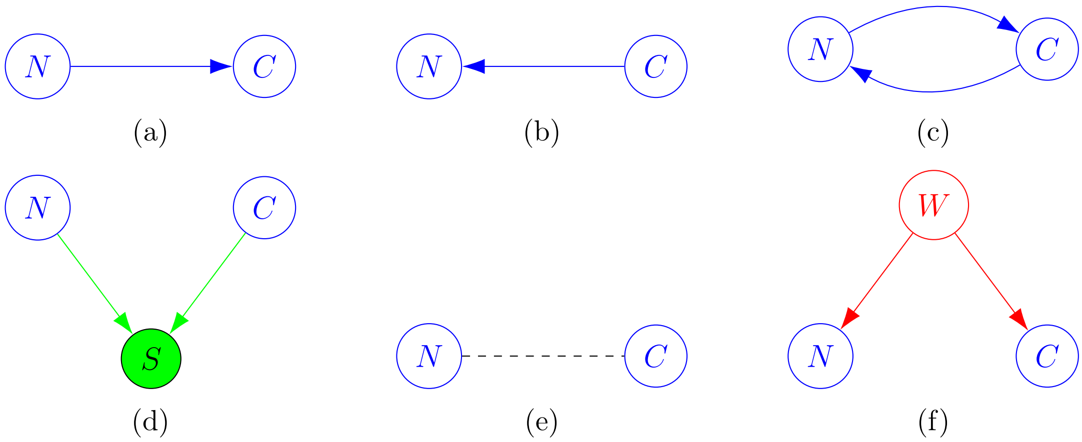


Figure 2: Graphical representations of different correlation inducing scenarios.

Discussion 1.1.2. *Correlation does not imply causation because there are other possible correlation inducing scenarios. Also, correlation is symmetric, causation is asymmetric.*

1.2. Causal Effects in the Real World

Example 1.2.1 (Does the thermometer cause the sun to rise?). *Consider an old type of thermometer (T) with a needle that can—for simplicity of arguments—either point to higher temperatures (up) or to lower temperatures (down). We also consider the state of the sun (S), which can either be up (u) or down (d). We then observe that T correlates with S. For simplicity, we assume a one-to-one relationship:*

T	S
u	u
d	d

The conditional distribution $P(S|T)$ then looks like this:

$$\begin{aligned} P(S=u|T=u) &= 1, \\ P(S=u|T=d) &= 0, \\ P(S=d|T=u) &= 0, \\ P(S=d|T=d) &= 1. \end{aligned}$$

If we are cold we are now tempted to try changing the needle in the thermometer in order to make the sun rise and warm us up.
What is wrong with our analysis?

Discussion 1.2.2. *The example 1.2.1 makes clear that there is a difference between:*

1. ***Observing** the positions of the thermometer needle T and the sun S, resulting in an **observational** data set, leading to an estimate of $P(S|T)$.*

2. ***Interacting** with the thermometer needle T and getting the sun's response S, resulting in an **interventional** data set, leading to estimates for $P(S|\operatorname{do}(T))$:*

$$\begin{aligned} P(S=u|\operatorname{do}(T=u)) &= 0.5, \\ P(S=u|\operatorname{do}(T=d)) &= 0.5, \\ P(S=d|\operatorname{do}(T=u)) &= 0.5, \\ P(S=d|\operatorname{do}(T=d)) &= 0.5. \end{aligned}$$

Definition 1.2.3 (Causal effect—informal definition)**.** *We say that a variable X has a **causal effect** on another variable Y if **forcing** X to take on a value x, the distribution of Y explicitly depends on x, that is:*

$$\exists x \in \mathcal{X}: \qquad P(Y|\operatorname{do}(X=x)) \neq P(Y).$$

Remark 1.2.4.

1. *Again, note that example 1.2.1 shows that the condition in definition 1.2.3 is different from:*

$$\exists x \in \mathcal{X}: \qquad P(Y|X=x) \neq P(Y).$$

which just uses the conditional distributions instead of the interventional distributions.

2. *Also note, that the* 'do-*operators' are not operators on the observational distribution $P(X,Y)$ or $P(Y|X)$, etc., or on the corresponding observational data sets. They reflect actions/interventions in the real world leading to different distributions and corresponding data sets.*

3. *There are usually many possible intervention values and targets one can think of, leading to many different interventional distributions and data sets.*

4. *One may think of the observational distribution as a special case of an interventional distribution (where we intervene by doing nothing).*

1.3. Randomized Controlled Trials (RCT)

Principle 1.3.1 (Randomized Controlled Trial (RCT))**.** *Assume we want to know if 'treatment' variable X has a causal effect on 'outcome' variable Y, i.e. we want to estimate the deviation between: $P(Y|\,\mathrm{do}(X = x_0))$ and $P(Y|\,\mathrm{do}(X = x_1))$. For this we have test subjects $w_1, \dots, w_N$. A* ***Randomized Controlled Trial*** *then follows the following steps:*

1. *Split the population of test subjects into 2 groups ('test group' C_1 vs. 'control group' C_0) by random lot (or fair coin flips).*

2. *Give every test subject $w_n \in C_1$ from 'test group' the treatment x_1 and the ones $w_n \in C_0$ from 'control group' the control treatment x_0.*

3. *Measure the outcome y_n for each test subject w_n and estimate the deviation:*

$$D := d(P(Y|\,\mathrm{do}(X = x_0)), P(Y|\,\mathrm{do}(X = x_1))).$$

4. *Do a statistical test if the deviation D is significantly different from 0.*

5. *If it is significantly different from 0 we can conclude a* ***causal effect*** *of X on Y, otherwise not.*

Remark 1.3.2. *The notion of a randomized controlled trial goes back several centuries. It was already described in 1648 by Flemish physician Jan Baptista van Helmont [vH48]: "Let us take from the itinerants' hospitals, from the camps or from elsewhere 200 or 500 poor people with fevers, pleurisy etc. and divide them in two: let us cast lots so that one half of them fall to me and the other half to you. I shall cure them without blood-letting or perceptible purging, you will do so according to your knowledge (nor do I even hold you to your boast of abstaining from phlebotomy or purging) and we shall see how many funerals each of us will have: the outcome of the contest shall be the reward of 300 florins deposited by each of us. Thus shall your business be concluded. O Magistrates to whose hearts the health of your people is dear; let the trial be made for the public good, in order to know the truth, for the sake of your life and soul and for the health of all the people, sons, widows and orphans. Let there be a real debate to find the means of cure."*

Example 1.3.3. *Example applications of randomized controlled trials are:*

1. *drug or vaccine testing,*

2. *advertisement placement,*

3. *evaluating public policies, etc.*

4. *A. Banerjee, E. Duflo, M. Kremer got the Nobel Prize in Economics 2019 for using RCTs in poverty research, e.g. improving school attendance and performance in poor areas via giving different towns different incentives (e.g. text books vs. deworming medicine vs. control groups).*

Discussion 1.3.4. *1. An RCT is an 'interventional study' (in contrast to just 'observational study') since we control the treatment and 'force' it onto the test subjects.*

2. Randomized Controlled Trials are considered the gold standard for experimental causal discovery.

3. To further avoid biases one usually insists on double/triple blind RCT studies, i.e. noone directly involved in the study knows who got which treatment (e.g. neither the doctor, the experimenter, the patient, etc.).

4. Often RCTs cannot be done for ethical reasons (e.g. "smoking causes cancer" research).

5. Sometimes RCTs require too many resources to be feasible.

Exercise 1.3.5. *Go online, find news like "drinking wine every day is good for your health" or "chewing gum causes diabetes", etc., look up the original research paper and check:*

1. *if they did interventional studies (like RCT) or just observational studies,*
2. *in case of an RCT, whether it was double/triple blind,*
3. *otherwise, if (and how) they ruled out other correlation inducing scenarios,*
4. *what bias could have possibly been introduced through the data collecting process,*
5. *how big the data set was, what assumptions were made, what statistical methods were used, etc.,*
6. *what other 'stories' you could come up with in order to explain the data.*

Write down your findings and talk to others about it.

2. Transition Probability Theory

2.1. Elementary Probability Theory

Example 2.1.1 (Winning a pie with a biased die)**.** *You are allowed to roll a biased die with 6 sides. If you roll a* 5 *or* 6 *you win a* ***car****, a* 4 *gives you a* ***mug*** *and* $1, 2, 3$ *wins you an apple* ***pie****. In this case the sample space is* $\mathcal{W} := \{1, 2, 3, 4, 5, 6\}$ *and the die introduces a probability distribution* P *on* $\mathcal{W}$*. Since the die is biased, we have to specify each of the probability masses to throw those numbers separately:*

$$p(1) = 0.5, \qquad p(2) = p(3) = p(4) = 0.1, \qquad p(5) = 0.15, \qquad p(6) = 0.05.$$

We are now interested in the probabilities of the events of winning those 3 different prices. For this we consider the 'prize' space: $\mathcal{Z} := \{\text{pie}, \text{mug}, \text{car}\}$*. We can then formalize the outcome via the map* F*:*

$$\begin{array}{rccc} F: & \mathcal{W} & \to & \mathcal{Z}, \\ & 1, 2, 3 & \mapsto & \text{pie}, \\ & 4 & \mapsto & \text{mug}, \\ & 5, 6 & \mapsto & \text{car}. \end{array}$$

To compute the probability of winning each of the prizes we need to 'push' the probability distribution P*, which lives on the space* $\mathcal{W}$*, to the space* $\mathcal{Z}$*. We can do this as follows:*

$$\begin{aligned} P(F = \text{pie}) = P(F^{-1}(\{\text{pie}\})) &= P(\{1, 2, 3\}) &&= p(1) + p(2) + p(3) &&= 0.7, \\ P(F = \text{mug}) = P(F^{-1}(\{\text{mug}\})) &= P(\{4\}) &&= p(4) &&= 0.1, \\ P(F = \text{car}) = P(F^{-1}(\{\text{car}\})) &= P(\{5, 6\}) &&= p(5) + p(6) &&= 0.2, \end{aligned}$$

where $F^{-1}(C) := \{w \in \mathcal{W} \,|\, F(w) \in C\}$ *is the* ***pre-image*** *of* $C \subseteq \mathcal{Z}$*.*

Discussion 2.1.2. *The simple example 2.1.1 already provides us with the main examples for the typical probability-theoretic terminology and important insights:*

1. *We call the tuple* $(\mathcal{W}, P)$ *a* ***probability space****. It is important to note that* P *was defined on* $\mathcal{W}$*, not* $\mathcal{Z}$*.*

2. *We call the map* F *a* ***random variable****, which is really nothing else than a map from a probability space to another space.*

3. ***Events*** *are modelled by* ***subsets*** $B \subseteq \mathcal{W}$*, not just by single elements* $w \in \mathcal{W}$*. For example consider the event that you don't win a car. This event can't be represented by a single element in* $\mathcal{W}$ *or* $\mathcal{Z}$*.*

4. *In this example we can compute the probability of an event by* ***additivity*** *of* P *and the use of the probability mass function, via* $P(B) = \sum_{w \in B} p(w)$*.*

5. *The distribution of the prizes, i.e. the distribution of random variable* F*, assigns probabilities to events* $C \subseteq \mathcal{Z}$ *and can be computed using the* ***pre-image*** *of* F *via* $P(F \in C) = P(F^{-1}(C))$*, where the latter is now an event* $F^{-1}(C) \subseteq \mathcal{W}$*, which we already know how to deal with.*

6. *The distribution of F on $\mathcal{Z}$ here is also called the* ***push-forward*** *distribution or* ***image*** *distribution of P via F or just the* ***law*** *of F. It is often abbreviated as: P_F, P^F, F_*P or $P(F)$. Again note: $P(F)(C) := P(F \in C) = P(F^{-1}(C))$.*

7. *So $(\mathcal{Z}, P(F))$ forms a probability space on its own and as soon as we know $P(F)$ we don't need any information about $(\mathcal{W}, P)$ anymore if all we are interested in is the events in $\mathcal{Z}$ and the law of F. All randomness on $\mathcal{Z}$ is fully specified by $P(F)$.*

Example 2.1.3. *Now consider the standard normal distribution $\mathcal{N}(0,1)$ on $\mathbb{R}$, which is specified by the* ***probability density*** *function:*

$$p(w) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \cdot w^2\right).$$

The probability of an event $A \subseteq \mathbb{R}$ is then given by:

$$P(A) = \int_A p(w)\, dw,$$

in case A can be integrated over (i.e. if it is not a too pathological set). For instance, if $A = [a,b] \cup [c,d]$ with $a \leq b < c \leq d$ we get:

$$P(A) = \int_a^b p(w)\, dw + \int_c^d p(w)\, dw.$$

Note that, even though $p(w) > 0$ for every $w \in \mathbb{R}$, we have:

$$P(\{x\}) = 0 \qquad \text{for every} \qquad x \in \mathbb{R}.$$

Now consider the random variable $F : \mathbb{R} \to \mathbb{R}$ with $F(w) = \sin(w)$. It is not immediately clear how to define the probability distribution of F when only working with probability densities. It is even more difficult to derive the probability density for F in this setting.

Discussion 2.1.4.

1. *The examples 2.1.1 and 2.1.3 show that many probability distributions can be represented either by* ***probability mass*** *functions (****discrete*** *case), $w \mapsto p(w)$, or* ***probability density*** *functions (****absolute continuous*** *case), $w \mapsto p(w)$.*

2. *Both cases have in common that one only needs a function that takes elements $w \in \mathcal{W}$ as arguments, in contrast to subsets $A \subseteq \mathcal{W}$. This is usually the reason why only the discrete and absolute continuous cases are taught in elementary probability theory or machine learning classes.*

3. *Note that, in the discrete case with K classes, one only needs to specify the K values $p(1), \dots, p(K)$, in contrast to the 2^K values on subsets $P(A)$ for $A \in 2^{\mathcal{W}}$ (the* ***power set*** *of $\mathcal{W}$ consisting of all subsets $A \subseteq \mathcal{W}$), as the latter values can be derived from the former values using additivity.*

4. *We have problems defining probability distributions of random variables for absolute continuous distributions when we are only allowed to work with probability densities.*

5. *Measure theory is the framework that directly works with subsets $A \subseteq \mathcal{W}$, in contrast to elements $w \in \mathcal{W}$, and provides a unifying language that encompasses both special cases.*

2.2. Recap - Measure Theoretic Probability

Here we just remind the reader of our notations for the core concepts of measure theoretic probability. More can be found in Appendix A.

2.2.1. Measurable Spaces and Maps

Definition 2.2.1 (σ-algebras). *Let $\mathcal{W}$ be a set. A (non-empty) set $\mathcal{B} \subseteq 2^{\mathcal{W}}$ of subsets $A \subseteq \mathcal{W}$ is called a* ***σ-algebra*** *on $\mathcal{W}$ if it satisfies the following rules:*

i) empty set: $\emptyset \in \mathcal{B}$,

ii) complement: If $A \in \mathcal{B}$ then also: $A^{\mathsf{c}} := \mathcal{W} \setminus A \in \mathcal{B}$,

iii) countable union: If $A_n \in \mathcal{B}$ for all $n \in \mathbb{N}$ then also: $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{B}$.

Definition 2.2.2 (Measurable spaces). *A tuple $(\mathcal{W}, \mathcal{B})$ of a set $\mathcal{W}$ and a σ-algebra $\mathcal{B}$ on $\mathcal{W}$ is called* ***measurable space****.*

Remark 2.2.3 (Abuse of notation). *By abuse of notation we often just call $\mathcal{W}$ a measurable space by implicitly assuming that it is endowed with a fixed σ-algebra, which we will indicate by $\mathcal{B}_{\mathcal{W}}$ or $\mathcal{B}(\mathcal{W})$ if needed. We will also just call a subset $A \subseteq \mathcal{W}$* ***measurable*** *when we actually mean that $A \in \mathcal{B}_{\mathcal{W}}$.*

Definition 2.2.4 (Measurable maps). *Let $(\mathcal{W}, \mathcal{B}_{\mathcal{W}})$ and $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$ be two measurable spaces and $f : \mathcal{W} \to \mathcal{Z}$ be a map. We call f a* ***$\mathcal{B}_{\mathcal{W}}$-$\mathcal{B}_{\mathcal{Z}}$-measurable map*** *(or just* ***measurable*** *for short) if for all $B \in \mathcal{B}_{\mathcal{Z}}$ the pre-image $f^{-1}(B)$ is an element of $\mathcal{B}_{\mathcal{W}}$. In formulas:*

$$\forall B \in \mathcal{B}_{\mathcal{Z}} : \; f^{-1}(B) \in \mathcal{B}_{\mathcal{W}}.$$

Remember the definition of pre-image: $f^{-1}(B) := \{w \in \mathcal{W} \,|\, f(w) \in B\}$.

For most of the lecture we will restrict to well-behaved measurable spaces, namely standard measurable spaces. The key point is that they all behave like the space $[0, 1]$, or $\mathbb{R}$, with its Borel-σ-algebra. So (almost) all results for $[0, 1]$ immediately translate to standard measurable spaces.

Definition 2.2.5 (Standard measurable space, see [Fre15] 424A-G). *A measurable space $(\mathcal{W}, \mathcal{B}_{\mathcal{W}})$ is called* ***standard measurable space*** *(aka standard Borel space) if it is measurably isomorphic to either:*

1. *a finite measurable space* $\{1, \dots, M\}$ *for some* $M \in \mathbb{N}$ *endowed with the power set* σ*-algebra* $2^{\{1,\dots,M\}}$*, or:*

2. *the countably infinite space* $\mathbb{N}$ *endowed with the power set* σ*-algebra* $2^{\mathbb{N}}$*, or:*

3. *the unit interval* $[0, 1]$ *endowed with its Borel* σ*-algebra*[2]*:*

$$\mathcal{B}_{[0,1]} = \sigma\left(\{[a, b] \mid a, b \in [0, 1] \cap \mathbb{Q}, a \leq b\}\right).$$

"Measurably isomorphic" means that there is a measurable map from one space to the other that has a measurable inverse.

The following theorem shows that (almost) all spaces we encounter in practice are actually standard measurable spaces, justifying our focus on standard measurable spaces for the most of this lecture.

Theorem 2.2.6 (Kuratowski et al., see [Fre15] 424A-G)**.** *Every Borel subset of any complete metric space that has a countable dense subset is a standard measurable space in its Borel* σ*-algebra.*

Example 2.2.7. $\mathbb{R}$, $\mathbb{R}^D$, $\mathbb{Q}$, $\mathbb{Z}$, $\mathbb{N}$, $\{1, \dots, M\}$, $[0, 1]$, *topological manifolds, countable CW-complexes, etc., are all standard measurable spaces.*

2.2.2. Finite and Probability Measures

Definition 2.2.8 (Measures)**.** *Let* $(\mathcal{W}, \mathcal{B})$ *be a measurable space. A* ***measure*** μ *on* $(\mathcal{W}, \mathcal{B})$*—by definition—is a map:*

$$\mu : \mathcal{B} \to \mathbb{R} \cup \{\infty\}, \quad D \mapsto \mu(D),$$

such that:

i) non-negative: $\forall A \in \mathcal{B}$*:* $\mu(A) \in [0, \infty]$,

ii) empty set: $\mu(\emptyset) = 0$,

iii) countably additive (aka σ*-additive): for all sequences* $A_n \in \mathcal{B}$*,* $n \in \mathbb{N}$*, with* $A_i \cap A_j = \emptyset$ *for all* $i \neq j$*, we have:*

$$\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

Definition 2.2.9 (Probability and finite measures)**.** *A measure* μ *on* $(\mathcal{W}, \mathcal{B}_{\mathcal{W}})$ *is called:*

1. ***probability measure*** *if* $\mu(\mathcal{W}) = 1$.

---

[2] See Definition A.3.2 for the σ-algebra generated by a set of subsets.

2. ***finite measure*** *if* $\mu(\mathcal{W}) < \infty$.

3. σ***-finite measure*** *if there are* $D_n \in \mathcal{B}$, $n \in \mathbb{N}$, *with* $\mu(D_n) < \infty$ *and* $\mathcal{W} = \bigcup_{n\in\mathbb{N}} D_n$.

Clearly, every probability measure is finite, and, every finite measure is σ-finite.

Definition 2.2.10 (The spaces of finite and probability measures)**.** *The set of all probability measures on* $(\mathcal{W}, \mathcal{B}_{\mathcal{W}})$ *is denoted by* $\mathcal{P}(\mathcal{W}, \mathcal{B}_{\mathcal{W}})$, *and the set of all finite measures by* $\mathcal{M}(\mathcal{W}, \mathcal{B}_{\mathcal{W}})$, *or* $\mathcal{P}(\mathcal{W})$ *and* $\mathcal{M}(\mathcal{W})$, *resp., for short. For* $B \in \mathcal{B}_{\mathcal{W}}$ *we consider the evaluation map:*

$$\mathrm{ev}_B : \mathcal{M}(\mathcal{W}) \to \mathbb{R}_{\geq 0}, \qquad \mu \mapsto \mathrm{ev}_B(\mu) := \mu(B).$$

We then endow $\mathcal{M}(\mathcal{W})$, *and* $\mathcal{P}(\mathcal{W})$, *resp., with the smallest* σ*-algebra* $\mathcal{B}$ *such that all evaluation maps* ev_B *are* $\mathcal{B}$-$\mathcal{B}_{\mathbb{R}_{\geq 0}}$*-measurable, where* $\mathcal{B}_{\mathbb{R}_{\geq 0}}$ *is the Borel-*σ*-algebra of* $\mathbb{R}_{\geq 0}$, *i.e.:*

$$\mathcal{B}_{\mathcal{M}(\mathcal{W})} := \sigma\left(\left\{\mathrm{ev}_B^{-1}((r,\infty)) \,\middle|\, B \in \mathcal{B}, r \in \mathbb{R}_{\geq 0}\right\}\right).$$

Remark 2.2.11. *The above definition implies that for measurable spaces* $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$, *a map:*

$$K : \mathcal{X} \to \mathcal{M}(\mathcal{Y}),$$

is $\mathcal{B}_{\mathcal{X}}$-$\mathcal{B}_{\mathcal{M}(\mathcal{Y})}$*-measurable if and only if for all* $B \in \mathcal{B}_{\mathcal{Y}}$ *the composition:*

$$\mathrm{ev}_B \circ K : \mathcal{X} \to \mathcal{M}(\mathcal{Y}) \to \mathbb{R}_{\geq 0},$$

is $\mathcal{B}_{\mathcal{X}}$-$\mathcal{B}_{\mathbb{R}_{\geq 0}}$*-measurable. Similarly, for* $\mathcal{P}(\mathcal{Y})$.

Theorem 2.2.12 (See [Par05] Thm. 6.2 + 6.5 or [Fre15] 437R)**.** *If* $(\mathcal{W}, \mathcal{B}_{\mathcal{W}})$ *is a standard measurable space then also* $\mathcal{P}(\mathcal{W})$ *is a standard measurable space (in its usual* σ*-algebra).*

2.2.3. The Measure Integral

For a measure μ on a measurable space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ the measure integral of measurable functions $f : \mathcal{X} \to \mathbb{R}$ is treated in Appendix A.5. Here we just want to remind the reader of our several different notations, which we will use interchangably during the course:

Notation 2.2.13 (Measure integral)**.** *We abbreviate the* ***measure integral*** *of a measurable function* $f : \mathcal{X} \to \mathbb{R}$ *w.r.t. measure* μ *on* $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ *as:*

$$\int f\, d\mu = \int f(x)\, d\mu(x) = \int f(x)\, \mu(dx).$$

If P *is a probability measure on* $\mathcal{X} = \mathbb{R}^D$ *that is either discrete or absolute continuous we have:*

$$\int f(x)\, P(dx) = \begin{cases} \sum_{x\in\mathcal{X}} f(x) \cdot p(x), & \text{if} \quad P \text{ is discrete,} \\ \int_{\mathcal{X}} f(x) \cdot p(x)\, dx & \text{if} \quad P \text{ is absolute continuous,} \end{cases}$$

where p *either denotes the probability mass function or the probability density, resp.*

2.2.4. The Lebesgue Measure

By far the most important measure is the Lebesgue measure, which assigns the typical D-dimensional volume to cubes, i.e. the product of their side lengths.

Definition 2.2.14 (The Lebesgue (outer) measure)**.** *The* ***Lebesgue (outer) measure*** *λ^D on $\mathbb{R}^D$ is given for subsets $A \subseteq \mathbb{R}^D$ via:*

$$\lambda^D(A) := \inf \left\{ \sum_{n \in \mathbb{N}} \mathrm{vol}^D \left([a^{(n)}, b^{(n)}] \right) \,\middle|\, A \subseteq \bigcup_{n \in \mathbb{N}} [a^{(n)}, b^{(n)}] \right\},$$

where the infimum is running over sequences of D-dimensional cubes:

$$[a^{(n)}, b^{(n)}] = [a_1^{(n)}, b_1^{(n)}] \times \cdots \times [a_D^{(n)}, b_D^{(n)}],$$

with $a^{(n)} = (a_1^{(n)}, \dots, a_D^{(n)})$, $b^{(n)} = (b_1^{(n)}, \dots, b_D^{(n)}) \in \mathbb{R}^D$, $a_d^{(n)} \leq b_d^{(n)}$ for $d = 1, \dots, D$, $n \in \mathbb{N}$, that jointly cover A, where the D-dimensional volume is given by:

$$\mathrm{vol}^D \left([a^{(n)}, b^{(n)}] \right) := (b_1^{(n)} - a_1^{(n)}) \cdots (b_D^{(n)} - a_D^{(n)}), \qquad \mathrm{vol}^D(\emptyset) := 0.$$

Theorem 2.2.15 (The Lebesgue measure)**.** *The* ***Lebesgue measure*** *λ^D, when restricted to the Borel-σ-algebra of $\mathbb{R}^D$, is the unique measure on $\mathbb{R}^D$ that satisfies:*

$$\lambda^D([a, b]) = \mathrm{vol}^D([a, b]),$$

for all D-dimensional cubes $[a, b]$. If the dimension is clear from the context we might just write λ for λ^D.

2.3. Finite Transition Measures and Markov Kernels

2.3.1. Core Definitions

Motivation 2.3.1. *If we consider a deterministic measurable map $f : \mathcal{T} \to \mathcal{W}$ then f assigns to each point $t \in \mathcal{T}$ exactly one point $w = f(t) \in \mathcal{W}$. Sometimes we rather want to model a* ***probabilistic map****, i.e. an assignment that can be random or comes with some uncertainties but still changes depending on the input t. The notion of* ***Markov kernels*** *formalizes this. A Markov kernel K from $\mathcal{T}$ to $\mathcal{W}$ can be considered a measurable map from $\mathcal{T}$ to the space of probability measures $\mathcal{P}(\mathcal{W})$ of $\mathcal{W}$:*

$$\mathcal{T} \to \mathcal{P}(\mathcal{W}).$$

It assigns to each $t \in \mathcal{T}$ a probability distribution over $\mathcal{W}$, which then assigns to each measurable subset $D \subseteq \mathcal{W}$ a probability value in $[0, 1]$.

Example 2.3.2 (Markov kernels)**.**

1. *any statistical model (i.e. family of model distributions) $\{p_\theta \,|\, \theta \in \mathcal{F}\}$, can be considered a Markov kernel, which we write $P(X|\Theta)$.*
2. *any conditional distribution $P(Y|X)$ can be considered a Markov kernel.*

3. *a neural network with softmax output for classification with input $x \in \mathcal{X}$, output $y \in \mathcal{Y}$ and weights $w \in \mathcal{W}$ can be seen as a Markov kernel $P(Y|X, W)$.*

We first start slightly more generally by defining finite transition measures.

Definition 2.3.3 (Finite transition measures and Markov kernels)**.** *Let $\mathcal{T}$, $\mathcal{W}$ be measurable spaces.*

1. *A **(finite[3]) transition measure** from $\mathcal{T}$ to $\mathcal{W}$ is—per definition—a measurable map:*
$$K : \mathcal{T} \to \mathcal{M}(\mathcal{W}),$$
from $\mathcal{T}$ to the space of finite measures of $\mathcal{W}$.

2. *A **transition probability** or **Markov kernel** is—per definition—a measurable map:*
$$K : \mathcal{T} \to \mathcal{P}(\mathcal{W}),$$
from $\mathcal{T}$ to the space of probability measures of $\mathcal{W}$.

Notation 2.3.4 (Transition measures and Markov kernels)**.**

1. *We often use suggestive notations as follows for finite transition measures and Markov kernels:*
$$K(W|T) : \mathcal{T} \to \mathcal{M}(\mathcal{W}), \qquad t \mapsto K(W|T = t),$$
where for every fixed $t \in \mathcal{T}$ the following map:
$$K(W|T = t) : \mathcal{B}_{\mathcal{W}} \to \mathbb{R}_{\geq 0}, \qquad D \mapsto K(W \in D|T = t),$$
is a finite measure, or probability measure, respectively.

2. *For fixed $D \in \mathcal{B}_{\mathcal{W}}$ we then use the following notation for the following measurable map:*
$$K(W \in D|T) : \mathcal{T} \to \mathbb{R}_{\geq 0}, \qquad t \mapsto K(W \in D|T = t).$$

3. *Since $K(W|T)$ takes the argument $t \in \mathcal{T}$ first, but then also $D \in \mathcal{B}_{\mathcal{W}}$ as a second argument we can also indentify $K(W|T)$ with the following two-argument map, which we denote with the same symbols:*
$$K(W|T) : \mathcal{B}_{\mathcal{W}} \times \mathcal{T} \to \mathbb{R}_{\geq 0}, \qquad (D, t) \mapsto K(W \in D|T = t).$$

4. *For Markov kernels $K(W|T)$ we will most of the time use the dashed arrow to $\mathcal{W}$ (instead of a usual arrow to $\mathcal{P}(\mathcal{W})$) to indicate the Markov kernel as follows:*
$$K(W|T) : \mathcal{T} \dashrightarrow \mathcal{W}, \quad (D, t) \mapsto K(W \in D|T = t).$$

[3] In this course we will only discuss *finite* transition measures.

5. *Note that above W and T are considered suggestive symbols only, but one could give W the meaning to mean the (identity or) projection map $\mathrm{pr}_{\mathcal{W}}$ onto $\mathcal{W}$. From the point on we also have a map T mapping to $\mathcal{T}$ the notation becomes ambiguous: $K(W|T)$ could also mean $K(W|T)$ where we plugged in T for t in "$T = t$", similar to conditional expectations $\mathbb{E}[W|T]$, but the meaning should become clear from the context.*

The implicit correspondence in the above discussion can more formally be summarized as:

Lemma 2.3.5. *There is a one-to-one correspondence between the following constructions:*

1. *a finite transition measure, i.e. a measurable map:*

$$K(W|T) : \mathcal{T} \to \mathcal{M}(\mathcal{W}), \qquad t \mapsto K(W|T=t).$$

2. *a two-argument function:*

$$\tilde{K}(W|T) : \mathcal{B}_{\mathcal{W}} \times \mathcal{T} \to \mathbb{R}_{\geq 0}, \qquad (D,t) \mapsto \tilde{K}(W \in D|T=t),$$

   such that:

   i) For each $t \in \mathcal{T}$ the map:

$$\mathcal{B}_{\mathcal{W}} \to \mathbb{R}_{\geq 0}, \qquad D \mapsto \tilde{K}(W \in D|T=t)$$

   is a finite measure (i.e. countably additive with $\tilde{K}(W \in \mathcal{W}|T=t) < \infty$ for all $t \in \mathcal{T}$).[4]

   ii) For each $D \in \mathcal{B}_{\mathcal{W}}$ the map:

$$\mathcal{T} \to \mathbb{R}_{\geq 0}, \quad t \mapsto \tilde{K}(W \in D|T=t)$$

   is $\mathcal{B}_{\mathcal{T}}$-$\mathcal{B}_{\mathbb{R}_{\geq 0}}$-measurable.

For Markov kernels the same statement holds after replacing $\mathcal{M}(\mathcal{W})$ with $\mathcal{P}(\mathcal{W})$ and "finite measure" with "probability measure".

Proof. The correspondence is via putting $K(W \in D|T=t) = \tilde{K}(W \in D|T=t)$ and vice versa. The corresponding properties hold by definition of the σ-algebra on $\mathcal{M}(\mathcal{W})$, also see Remark 2.2.11. Working out the details is left as an exercise. $\square$

[4]Note that for a finite transition measure the finite value $\tilde{K}(W \in \mathcal{W}|T=t)$ can vary with $t \in \mathcal{T}$. This is in contrast to Markov kernels where we always have $\tilde{K}(W \in \mathcal{W}|T=t) = 1$ for all $t \in \mathcal{T}$.

2.3.2. Special Cases of Markov Kernels

Example 2.3.6 (Markov kernels on discrete spaces)**.** *Consider a Markov kernel:*

$$K(W|T): \mathcal{T} \dashrightarrow \mathcal{W}, \quad (D,t) \mapsto K(W \in D|T=t),$$

where both $\mathcal{W} = \{w_1, \dots, w_M\}$ and $\mathcal{T} = \{t_1, \dots, t_K\}$ are finite discrete spaces. Then we can define the mass function k via:

$$k(w_i|t_j) := K(W \in \{w_i\}|T = t_j),$$

and the matrix $\tilde{K} := (k(w_i|t_j))_{i,j}$. Then the matrix $\tilde{K}$ is a stochastic matrix, i.e. it has non-negative entries and each of its columns sums to 1. $\tilde{K}$ then fully determines the Markov kernel K. So in the (finite) discrete case a Markov kernel is basically nothing else than a stochastic matrix filled with the transition probabilities.

Example 2.3.7 (Linear Gaussian Markov kernels)**.** *Let $\mathcal{W} = \mathbb{R}^M$, $\mathcal{T} = \mathbb{R}^L$, $\gamma \in \mathbb{R}^M$, $\Gamma \in \mathbb{R}^{M \times L}$ and $\Sigma \in \mathbb{R}^{M \times M}$ a fixed symmetric, positive-definite covariance matrix. Then:*

$$K(W \in D|T=t) := \int_D \mathcal{N}(w|\Gamma \cdot t + \gamma, \Sigma)\, dw,$$

*defines a Markov kernel from $\mathcal{T}$ to $\mathcal{W}$. Markov kernels of this form are called **linear Gaussian Markov kernels**. If Σ is only positive-semi-definite we call $K(W|T)$ a **degenerate** or **generalized linear Gaussian Markov kernel**.*

Example 2.3.8 (Exponential families as finite transition measures)**.** *Let $\mathcal{W}$ be a measurable space and μ a (non-zero) measure on $\mathcal{W}$ and $S : \mathcal{W} \to \mathbb{R}^D$ a measurable map. Define for $t \in \mathbb{R}^D$:*

$$Z(t) := \int_{\mathcal{W}} \exp\left(t^\top S(w)\right) \mu(dw) \quad \in \quad (0, \infty].$$

We then put:

$$\mathcal{T} := \left\{ t \in \mathbb{R}^D \,\middle|\, Z(t) < \infty \right\}.$$

We can then define the finite transition measure $K(W|T)$ from $\mathcal{T}$ to $\mathcal{W}$ for $D \in \mathcal{B}_{\mathcal{W}}$ and $t \in \mathcal{T}$ via:

$$K(W \in D|T=t) := \int_D \exp\left(t^\top S(w)\right) \mu(dw).$$

From this we get the Markov kernel $Q(W|T): \mathcal{T} \dashrightarrow \mathcal{W}$ via normalization:

$$Q(W \in D|T=t) := \int_D \exp\left(t^\top S(w) - L(t)\right) \mu(dw),$$

with log-normalizer: $L(t) := \log Z(t) = \log K(W \in \mathcal{W}|T=t)$.

Remark 2.3.9 (Markov kernels generalize probability distributions)**.** *Let $\mathcal{W}$ be a measurable space.*

1. *Every probability distribution $P \in \mathcal{P}(\mathcal{W})$ can be considered as a* ***constant Markov kernel*** *from $\mathcal{T}$ to $\mathcal{W}$ via:*

$$K : \mathcal{T} \dashrightarrow \mathcal{W}, \quad (D, t) \mapsto K(D|t) := P(D).$$

2. *Every Markov kernel from the one-point space: $\mathcal{T} = * := \{*\}$ to $\mathcal{W}$:*

$$K : * \dashrightarrow \mathcal{W}, \quad (D, *) \mapsto K(D|*),$$

defines a unique probability distribution $P \in \mathcal{P}(\mathcal{W})$ given via:

$$P(D) := K(D|*).$$

So we can identify probability distributions on $\mathcal{W}$ with Markov kernels $* \dashrightarrow \mathcal{W}$.

Remark 2.3.10 (Markov kernels generalize deterministic maps)**.** *Consider a measurable mapping $f : \mathcal{T} \to \mathcal{W}$. Then we can turn f into a Markov kernel δ_f via:*

$$\delta_f : \mathcal{T} \dashrightarrow \mathcal{W}, \quad (D, t) \mapsto \delta_f(D|t) := \mathbb{1}_D(f(t)),$$

which puts 100% probability mass onto the function value $f(t)$ for given $t \in \mathcal{T}$.

2.3.3. The Doob-Radon-Nikodym Derivative

Definition 2.3.11 (Absolute continuity)**.** *Let $\mathcal{T}$, $\mathcal{W}$ be measurable spaces and*

$$Q(W|T),\, K(W|T) : \mathcal{T} \to \mathcal{M}(\mathcal{W}),$$

two finite transition measures. We say that $Q(W|T)$ is ***absolute continuous w.r.t.*** *$K(W|T)$ if for all $t \in \mathcal{T}$ and $D \in \mathcal{B}_\mathcal{W}$ we have the implication:*

$$K(W \in D|T = t) = 0 \quad \implies \quad Q(W \in D|T = t) = 0.$$

In symbols we abbreviate this as:

$$Q(W|T) \ll K(W|T).$$

Remark 2.3.12. *For absolute continuous finite transition measures $Q(W|T) \ll K(W|T)$ there exists by the Theorem of Radon-Nikodym, see Theorem A.6.4 or [Kle20] Cor. 7.34, for each $t \in \mathcal{T}$ separately a Radon-Nikodym derivative, i.e. a $\mathcal{B}_\mathcal{W}$-$\mathcal{B}_{\mathbb{R}_{\geq 0}}$-measurable map:*

$$g_t : \mathcal{W} \to \mathbb{R}_{\geq 0},$$

such that for all $D \in \mathcal{B}_\mathcal{W}$:

$$Q(W \in D|T = t) = \int \mathbb{1}_D(w) \cdot g_t(w)\, K(W \in dw|T = t).$$

Unfortunately, the map:

$$g : \mathcal{W} \times \mathcal{T} \to \mathbb{R}_{\geq 0}, \qquad g(w|t) := g_t(w),$$

is not guaranteed to be jointly measurable, i.e. $(\mathcal{B}_{\mathcal{W}} \otimes \mathcal{B}_{\mathcal{T}})$-$\mathcal{B}_{\mathbb{R}_{\geq 0}}$*-measurable. In case it was, we would call it a* ***Doob-Radon-Nikodym derivative*** *of* $Q(W|T)$ *w.r.t.* $K(W|T)$*. Doob invented an alternative, but a bit more restrictive approach than the usual one to construct Radon-Nikodym derivatives for measures based on martingales. This approach will be seen to also work for the construction of Doob-Radon-Nikodym derivatives for finite transition measures.*

Definition 2.3.13 (Doob-Radon-Nikodym derivative)**.** *Let* $\mathcal{T}$*,* $\mathcal{W}$ *be measurable spaces and*

$$Q(W|T),\, K(W|T) : \mathcal{T} \to \mathcal{M}(\mathcal{W}),$$

two finite transition measures. A map

$$g : \mathcal{W} \times \mathcal{T} \to \mathbb{R}_{\geq 0}, \qquad (w,t) \mapsto g(w|t),$$

is called ***Doob-Radon-Nikodym derivative*** *if* g *is* $(\mathcal{B}_{\mathcal{W}} \otimes \mathcal{B}_{\mathcal{T}})$-$\mathcal{B}_{\mathbb{R}_{\geq 0}}$*-measurable and for all* $t \in \mathcal{T}$ *and all* $D \in \mathcal{B}_{\mathcal{W}}$ *we have:*

$$Q(W \in D|T = t) = \int \mathbb{1}_D(w) \cdot g(w|t)\, K(W \in dw|T = t).$$

In other words, g *provides a Radon-Nikodym derivative simultaneously for all* $t \in \mathcal{T}$*:*

$$g(w|t) = \frac{Q(W \in dw|T = t)}{K(W \in dw|T = t)}(w),$$

that is even jointly measurable in (w, t)*.*

Lemma 2.3.14. *If* $Q(W|T)$ *has a Doob-Radon-Nikodym derivative w.r.t.* $K(W|T)$ *then* $Q(W|T)$ *is absolute continuous w.r.t.* $K(W|T)$*.*

Proof. Should be clear, left as an exercise. □

To investigate the uniqueness of the Doob-Radon-Nikodym derivative we need the following notion of $K(W|T)$-null sets.

Definition 2.3.15 (Null sets)**.** *Let* $K(W|T) : \mathcal{T} \to \mathcal{M}(\mathcal{W})$ *be a finite transition measure. A subset* $N \subseteq \mathcal{W} \times \mathcal{T}$ *is called* $K(W|T)$***-null*** *if* $N_t := \{w \in \mathcal{W} \,|\, (w,t) \in N\}$ *is a* $K(W|T = t)$*-null set for every* $t \in \mathcal{T}$*, i.e. if for every* $t \in \mathcal{T}$ *there exists a measurable set* $M_t \in \mathcal{B}_{\mathcal{W}}$ *such that* $K(W \in M_t|T = t) = 0$ *and* $N_t \subseteq M_t$*.*

Lemma 2.3.16 (Essential uniqueness of the Doob-Radon-Nikodym derivative)**.** *Let* $\mathcal{T}$*,* $\mathcal{W}$ *be measurable spaces and:*

$$Q(W|T), K(W|T) : \mathcal{T} \to \mathcal{M}(\mathcal{W}),$$

be two finite transition measures with $Q(W|T) \ll K(W|T)$ and let g_1, g_2 be two Doob-Radon-Nikodym derivatives. Then the set:

$$N := \{(w,t) \in \mathcal{W} \times \mathcal{T} \,|\, g_1(w|t) \neq g_2(w|t)\}$$

is a $K(W|T)$-null set and an element of the product σ-algebra $\mathcal{B}_\mathcal{W} \otimes \mathcal{B}_\mathcal{T}$. In this sense, the Doob-Radon-Nikodym derivative is essentially unique.

Theorem 2.3.17 (Doob-Radon-Nikodym, see [DM83] Thm. 58, [Kle20] Ex. 11.17)**.** *Let $\mathcal{T}$, $\mathcal{W}$ be measurable spaces and:*

$$K(W|T),\, Q(W|T) : \mathcal{T} \to \mathcal{M}(\mathcal{W}),$$

be two finite transition measures. Assume that $\mathcal{W}$ is a standard measurable space.[5] Then the following two statements are equivalent:

1. *$Q(W|T)$ is absolute continuous w.r.t. $K(W|T)$.*
2. *$Q(W|T)$ has a Doob-Radon-Nikodym derivative w.r.t. $K(W|T)$.*

In that case the Doob-Radon-Nikodym derivative is essentially unique.

Remark 2.3.18.

1. *As mentioned in the footnote[5] Theorem 2.3.17 still holds if one only requires $\mathcal{B}_\mathcal{W}$ to be countably generated. Further extensions could be made to σ-algebras $\mathcal{B}_\mathcal{W}$ that are countably generated up to some form of null-sets.*
2. *With more technical conditions one could extend Theorem 2.3.17 to work for σ-finite transition measures. A simple, but important, special case is treated in the following Corollary 2.3.19.*

Corollary 2.3.19 (Doob-Radon-Nikodym derivatives w.r.t. σ-finite measures)**.** *Let $\mathcal{T}$, $\mathcal{W}$ be measurable spaces, where $\mathcal{W}$ is a standard[5] measurable space, let*

$$P(W|T) : \mathcal{T} \to \mathcal{M}(\mathcal{W}),$$

be a finite transition measure and μ be a σ-finite measure on $\mathcal{W}$. Then the following two statements are equivalent:

1. *$P(W|T)$ is absolute continuous w.r.t. μ, i.e. for all $t \in \mathcal{T}$ and $D \in \mathcal{B}_\mathcal{W}$:*

$$\mu(D) = 0 \qquad \Longrightarrow \qquad P(W \in D|T = t) = 0.$$

2. *$P(W|T)$ has a Doob-Radon-Nikodym derivative w.r.t. μ, i.e. a jointly measurable map:*

$$p : \mathcal{W} \times \mathcal{T} \to \mathbb{R}_{\geq 0}, \qquad (w,t) \mapsto p(w|t),$$

such that for all $t \in \mathcal{T}$ and $D \in \mathcal{B}_\mathcal{W}$:

$$P(W \in D|T = t) = \int_D p(w|t)\, \mu(dw).$$

[5] The proof shows that we actually only require that $\mathcal{B}_\mathcal{W}$ is *countably generated.*

In that case the Doob-Radon-Nikodym derivative is essentially unique, i.e. for two such p, say p_1 and p_2, the set:

$$N := \{(w,t) \in \mathcal{W} \times \mathcal{T} \,|\, p_1(w|t) \neq p_2(w|t)\} \in \mathcal{B}_\mathcal{W} \otimes \mathcal{B}_\mathcal{T},$$

satisfies $\mu(N_t) = 0$ for all $t \in \mathcal{T}$.

Proof. Let $P(W|T)$ be absolute continuous w.r.t. μ. Since μ is σ-finite there exists a probability measure $Q(W)$ with:

$$Q(W) \ll \mu \ll Q(W).$$

Indeed, if μ is finite, we can just put $Q(W \in D) := \frac{\mu(D)}{\mu(\mathcal{W})}$. If μ is σ-finite, but not finite, then we have a decomposition $\mathcal{W} = \dot{\bigcup}_{n\in\mathbb{N}} \mathcal{W}_n$ with $0 < \mu(\mathcal{W}_n) < \infty$. We can then put:

$$Q(W \in D) := \sum_{n\in\mathbb{N}} 2^{-n} \frac{\mu(D \cap \mathcal{W}_n)}{\mu(\mathcal{W}_n)}.$$

By the standard Radon-Nikodym theorem there exists a Radon-Nikodym derivative q of $Q(W)$ w.r.t. μ. Note that $Q(W)$ defines the constant Markov kernel $Q(W|T)$ via $Q(W|T = t) := Q(W)$. We thus have the absolute continuity:

$$P(W|T) \ll \mu \ll Q(W|T).$$

By Theorem 2.3.17 we thus get a Doob-Radon-Nikodym derivative k of $P(W|T)$ w.r.t. $Q(W|T)$. Then p given by:

$$p(w|t) := k(w|t) \cdot q(w),$$

is a Doob-Radon-Nikodym derivative of $P(W|T)$ w.r.t. μ. Indeed, we get for all $t \in \mathcal{T}$ and $D \in \mathcal{B}_\mathcal{W}$:

$$P(W \in D|T = t) = \int_D k(w|t)\, Q(W \in dw|T = t) = \int_D k(w|t) \cdot q(w)\, \mu(dw).$$

This shows one direction.

The essential uniqueness follows similar to Lemma 2.3.16 and the other direction similar to Lemma 2.3.14. $\square$

Corollary 2.3.20 (Absolute continuity and strictly positive densities)**.** *Let $\mathcal{T}$, $\mathcal{W}$ be measurable spaces, where $\mathcal{W}$ is standard*[5]*, and:*

$$P(W|T),\, K(W|T),\, Q(W|T) : \mathcal{T} \to \mathcal{M}(\mathcal{W}),$$

be finite transition measures and μ be a σ-finite measure on $\mathcal{W}$.

1. *$Q(W|T)$ has a* ***strictly positive*** *Doob-Radon-Nikodym derivative w.r.t. $K(W|T)$ if and only if:*

$$Q(W|T) \ll K(W|T) \ll Q(W|T).$$

2. *$P(W|T)$ has a* ***strictly positive*** *Doob-Radon-Nikodym derivative w.r.t. μ if and only if:*
$$\mu \ll P(W|T) \ll \mu.$$

Proof. The second case follows from the first using the arguments from Corollary 2.3.19. So, first assume that $Q(W|T)$ has a strictly positive density $q > 0$ w.r.t. $K(W|T)$. Then by Lemma 2.3.14 we already have: $Q(W|T) \ll K(W|T)$. Since q is strictly positive we can put for $w \in \mathcal{W}$ and $t \in \mathcal{T}$:

$$k(w|t) := \frac{1}{q(w|t)} > 0.$$

Then k is a (strictly positive) density of $K(W|T)$ w.r.t. $Q(W|T)$ and we again can use Lemma 2.3.14 to also get: $K(W|T) \ll Q(W|T)$. This shows one direction.

Now assume that we have:

$$Q(W|T) \ll K(W|T) \ll Q(W|T).$$

Then by the Doob-Radon-Nikodym Theorem 2.3.17 we have Doob-Radon-Nikodym derivatives q and k of $Q(W|T)$ w.r.t. $K(W|T)$ and of $K(W|T)$ w.r.t. $Q(W|T)$, resp. For all $D \in \mathcal{B}_\mathcal{W}$ and $t \in \mathcal{T}$ we thus get:

$$\begin{aligned}\int_D 1\, Q(W \in dw|T = t) &= Q(W \in D|T = t)\\ &= \int_D q(w|t)\, K(W \in dw|T = t)\\ &= \int_D q(w|t) \cdot k(w|t)\, Q(W \in dw|T = t).\end{aligned}$$

Since this holds for all $D \in \mathcal{B}_\mathcal{W}$ and $t \in \mathcal{T}$ we get that the set:

$$N := \{(w, t) \in \mathcal{W} \times \mathcal{T} \,|\, 1 \neq q(w|t) \cdot k(w|t)\} \in \mathcal{B}_\mathcal{W} \otimes \mathcal{B}_\mathcal{T},$$

is a $Q(W|T)$-null set, and because $K(W|T) \ll Q(W|T)$, also a $K(W|T)$-null set. We then put:

$$\tilde{q}(w|t) := q(w|t) \cdot \mathbb{1}_{N^c}(w, t) + \mathbb{1}_N(w, t), \tag{1}$$
$$\tilde{k}(w|t) := k(w|t) \cdot \mathbb{1}_{N^c}(w, t) + \mathbb{1}_N(w, t). \tag{2}$$

These are then still corresponding Doob-Radon-Nikodym derivatives and satisfy for all $w \in \mathcal{W}$ and $t \in \mathcal{T}$:
$$\tilde{q}(w|t) \cdot \tilde{k}(w|t) = 1,$$
which directly implies: $\tilde{q}(w|t), \tilde{k}(w|t) > 0$ for all $w \in \mathcal{W}$ and $t \in \mathcal{T}$. This shows the claim. $\square$

Proofs - Theorem of Doob-Radon-Nikodym

Lemma 2.3.21 (Essential uniqueness of the Doob-Radon-Nikodym derivative)**.** *Let $\mathcal{T}$, $\mathcal{W}$ be measurable spaces and:*

$$Q(W|T), K(W|T): \mathcal{T} \to \mathcal{M}(\mathcal{W}),$$

be two finite transition measures with $Q(W|T) \ll K(W|T)$ and let g_1, g_2 be two Doob-Radon-Nikodym derivatives. Then the set:

$$N := \{(w,t) \in \mathcal{W} \times \mathcal{T} \,|\, g_1(w|t) \neq g_2(w|t)\}$$

is a $K(W|T)$-null set and an element of the product σ-algebra $\mathcal{B}_\mathcal{W} \otimes \mathcal{B}_\mathcal{T}$.

Proof. Consider the set:

$$N^> := \{(w,t) \in \mathcal{W} \times \mathcal{T} \,|\, g_1(w|t) > g_2(w|t)\} = (g_1 \times g_2)^{-1}(\Delta^>),$$

where $\Delta^>$ is the measurable set:

$$\Delta^> := \{(r_1, r_2) \in \mathbb{R} \times \mathbb{R} \,|\, r_1 > r_2\} \in \mathcal{B}_{\mathbb{R}^2}.$$

Since both g_1 and g_2 are jointly measurable that shows that $N^> \in \mathcal{B}_\mathcal{W} \otimes \mathcal{B}_\mathcal{T}$. It follows that $N_t^> \in \mathcal{B}_\mathcal{W}$. Furthermore, we get:

$$\begin{aligned}
0 &= Q(W \in N_t^>|T=t) - Q(W \in N_t^>|T=t) \\
&= \int \mathbb{1}_{N_t^>}(w) \cdot g_1(w|t)\, K(W \in dw|T=t) - \int \mathbb{1}_{N_t^>}(w) \cdot g_2(w|t)\, K(W \in dw|T=t) \\
&= \int \underbrace{\mathbb{1}_{N_t^>}(w) \cdot (g_1(w|t) - g_2(w|t))}_{>0 \text{ for } w \in N_t^>}\, K(W \in dw|T=t).
\end{aligned}$$

This shows that $K(W \in N_t^>|T=t) = 0$. By flipping g_1 and g_2 we also get: $K(W \in N_t^<|T=t) = 0$ and thus $K(W \in N_t|T=t) = 0$, where we notice that $N = N^> \cup N^<$. This shows the claim. ☐

Theorem 2.3.22 (Existence of the Doob-Radon-Nikodym derivative, see [DM83] Thm. 58, [Kle20] Ex. 11.17)**.** *Let $\mathcal{T}$, $\mathcal{W}$ be measurable spaces and:*

$$K(W|T),\, Q(W|T): \mathcal{T} \to \mathcal{M}(\mathcal{W}),$$

be two finite transition measures. Assume that $\mathcal{W}$ is a standard measurable space.[5] *$Q(W|T) \ll K(W|T)$ implies that $Q(W|T)$ has a Doob-Radon-Nikodym derivative w.r.t. $K(W|T)$.*

Proof sketch. Since $\mathcal{W}$ is a standard measurable space we have that $\mathcal{B}_\mathcal{W}$ is countably generated, i.e. $\mathcal{B}_\mathcal{W} = \sigma(\mathcal{S})$ with a countable $\mathcal{S} = \{D_n \,|\, n \in \mathbb{N}\} \subseteq \mathcal{B}_\mathcal{W}$. If for example, $\mathcal{W} = [0,1]$, which we can w.l.o.g. assume, then we could choose $\mathcal{S} = \{[a,b] \,|\, a \le b, a, b \in \mathbb{Q} \cap [0,1]\}$. We now define the following sequence of finite measurable partitions of $\mathcal{W}$ inductively via:

$$\mathcal{E}_0 := \{\mathcal{W}\}, \qquad \mathcal{E}_{n+1} := \left(\bigcup_{D \in \mathcal{E}_n} \{D \setminus D_n, D \cap D_n\} \right) \setminus \{\emptyset\}, \quad n \in \mathbb{N}.$$

We put $\mathcal{B}_n := \sigma(\mathcal{E}_n)$. Note that each $\mathcal{E}_n$ is finite and for every $n \in \mathbb{N}$:

$$\mathcal{W} = \dot{\bigcup_{D \in \mathcal{E}_n}} D, \qquad \mathcal{B}_n \subseteq \mathcal{B}_{n+1} \subseteq \mathcal{B}_\mathcal{W} = \sigma\left(\bigcup_{m \in \mathbb{N}} \mathcal{E}_m \right).$$

For $D \in \mathcal{B}_\mathcal{W}$ we can define the map $q_D : \mathcal{T} \to \mathbb{R}_{\ge 0}$ via:

$$q_D(t) := \frac{Q(W \in D|T = t)}{K(W \in D|T = t)} \cdot \mathbb{1}_{K(W \in D|T=t)>0} = \begin{cases} \frac{Q(W \in D|T=t)}{K(W \in D|T=t)}, & \text{if} \quad K(W \in D|T = t) > 0, \\ 0, & \text{if} \quad K(W \in D|T = t) = 0. \end{cases}$$

Since $Q(W \in D|T = t)$ and $K(W \in D|T = t)$ are measurable in t for each fixed D we see that q_D is $\mathcal{B}_\mathcal{T}$-$\mathcal{B}_{\mathbb{R}_{\ge 0}}$-measurable. For $n \in \mathbb{N}$ we now define:

$$G_n(w, t) := \sum_{D \in \mathcal{E}_n} \mathbb{1}_D(w) \cdot q_D(t),$$

and:

$$G(w, t) := \liminf_{n \in \mathbb{N}} G_n(w, t), \qquad g(w|t) := G(w, t) \cdot \mathbb{1}_{G(w,t)<\infty}.$$

We immediately see that every G_n is a $(\mathcal{B}_\mathcal{W} \otimes \mathcal{B}_\mathcal{T})$-$\mathcal{B}_{\mathbb{R}_{\ge 0}}$-measurable map. As a countable limit of measurable functions also G and g are $\mathcal{B}_\mathcal{W} \otimes \mathcal{B}_\mathcal{T}$-measurable. We claim that g is a Doob-Radon-Nikodym derivative of $Q(W|T)$ w.r.t. $K(W|T)$. Since we already showed that g is jointly measurable we are left to show that for every $t \in \mathcal{T}$ and $D \in \mathcal{B}_\mathcal{W}$ we have:

$$Q(W \in D|T = t) = \int \mathbb{1}_D(w) \cdot g(w|t) \, K(W \in dw|T = t).$$

So in the following we can fix $t \in \mathcal{T}$ and only indicate the dependence on t with an index:

$$G_n^t(w) := G_n(w, t), \qquad G^t(w) := G(w, t).$$

Notice that G_n^t is $\mathcal{B}_n$-measurable for $n \in \mathbb{N}$. In the following we will use that by construction of the $\mathcal{E}_n$ for $D \in \mathcal{E}_n$ and $m \ge n$ we have the disjoint union decompositions:

$$D = \dot{\bigcup_{\substack{A \in \mathcal{E}_m \\ A \subseteq D}}} A, \qquad \mathcal{W} = \dot{\bigcup_{D \in \mathcal{E}_n}} \left(\dot{\bigcup_{\substack{A \in \mathcal{E}_m \\ A \subseteq D}}} A \right).$$

Let $m \geq n$ then we get:

$$
\begin{aligned}
&G_n^t(w) \\
&= \sum_{D \in \mathcal{E}_n} \left[\frac{Q(W \in D|T=t)}{K(W \in D|T=t)} \cdot \mathbb{1}_{K(W \in D|T=t)>0} \right] \cdot \mathbb{1}_D(w) \\
&= \sum_{D \in \mathcal{E}_n} \left[\sum_{\substack{A \in \mathcal{E}_m \\ A \subseteq D}} \frac{Q(W \in A|T=t)}{K(W \in D|T=t)} \cdot \mathbb{1}_{K(W \in D|T=t)>0} \right] \cdot \mathbb{1}_D(w) \\
&= \sum_{D \in \mathcal{E}_n} \left[\sum_{\substack{A \in \mathcal{E}_m \\ A \subseteq D}} \frac{Q(W \in A|T=t)}{K(W \in D|T=t)} \cdot \mathbb{1}_{Q(W \in A|T=t)>0} \cdot \mathbb{1}_{K(W \in D|T=t)>0} \right] \cdot \mathbb{1}_D(w) \\
&\overset{Q \ll K}{=} \sum_{D \in \mathcal{E}_n} \left[\sum_{\substack{A \in \mathcal{E}_m \\ A \subseteq D}} \frac{Q(W \in A|T=t)}{K(W \in D|T=t)} \cdot \mathbb{1}_{K(W \in A|T=t)>0} \cdot \mathbb{1}_{K(W \in D|T=t)>0} \right] \cdot \mathbb{1}_D(w) \\
&= \sum_{D \in \mathcal{E}_n} \left[\sum_{\substack{A \in \mathcal{E}_m \\ A \subseteq D}} \frac{Q(W \in A|T=t)}{K(W \in A|T=t)} \cdot \frac{K(W \in A|T=t)}{K(W \in D|T=t)} \cdot \mathbb{1}_{K(W \in A|T=t)>0} \cdot \mathbb{1}_{K(W \in D|T=t)>0} \right] \cdot \mathbb{1}_D(w) \\
&= \sum_{A \in \mathcal{E}_m} \sum_{\substack{D \in \mathcal{E}_n \\ D \supseteq A}} \frac{Q(W \in A|T=t)}{K(W \in A|T=t)} \cdot \frac{K(W \in A|T=t)}{K(W \in D|T=t)} \cdot \mathbb{1}_{K(W \in A|T=t)>0} \cdot \mathbb{1}_{K(W \in D|T=t)>0} \cdot \mathbb{1}_D(w) \\
&= \sum_{A \in \mathcal{E}_m} \frac{Q(W \in A|T=t)}{K(W \in A|T=t)} \cdot \mathbb{1}_{K(W \in A|T=t)>0} \underbrace{\left[\sum_{\substack{D \in \mathcal{E}_n \\ D \supseteq A}} \cdot \frac{K(W \in A|T=t)}{K(W \in D|T=t)} \cdot \mathbb{1}_{K(W \in D|T=t)>0} \cdot \mathbb{1}_D(w) \right]}_{= \mathbb{E}_t[\mathbb{1}_A|\mathcal{B}_n](w)} \\
&= \sum_{A \in \mathcal{E}_m} \left[\frac{Q(W \in A|T=t)}{K(W \in A|T=t)} \cdot \mathbb{1}_{K(W \in A|T=t)>0} \right] \cdot \mathbb{E}_t[\mathbb{1}_A|\mathcal{B}_n](w) \\
&= \mathbb{E}_t \left[\sum_{A \in \mathcal{E}_m} \frac{Q(W \in A|T=t)}{K(W \in A|T=t)} \cdot \mathbb{1}_{K(W \in A|T=t)>0} \cdot \mathbb{1}_A \,\middle|\, \mathcal{B}_n \right] (w) \\
&= \mathbb{E}_t \left[G_m^t \,\middle|\, \mathcal{B}_n \right] (w).
\end{aligned}
$$

Note that we use $\mathbb{E}_t[\_|\mathcal{B}_n]$ to indicate conditional expectations w.r.t. $K(W|T=t)$ and $\mathcal{B}_n$. For the first conditional expectation see [Kle20] Lem. 8.10. So we get that G_n^t is a version of $\mathbb{E}_t [G_m^t \,|\, \mathcal{B}_n]$ for all $m \geq n$. This shows that $(G_n^t)_{n \in \mathbb{N}}$ is a martingale attached to the filtration $(\mathcal{B}_n)_{n \in \mathbb{N}}$ w.r.t. $K(W|T=t)$. Furthermore, we can show that $(G_n^t)_{n \in \mathbb{N}}$ is uniformly integrable w.r.t. $K(W|T=t)$, see [Kle20] Ex. 7.39. By the convergence theorem for uniformly integrable martingales, see [Kle20] Thm. 11.7, we get that G_n^t also converges in L^1 to G^t w.r.t. $K(W|T=t)$ and that G_n^t is a version of $\mathbb{E}_t[G^t|\mathcal{B}_n]$ for all $n \in \mathbb{N}$. So for $D \in \mathcal{E}_n$ the function $\mathbb{1}_D \cdot G_n^t$ is a version of $\mathbb{E}_t[\mathbb{1}_D \cdot G^t|\mathcal{B}_n]$. Taking expectation values shows:

$$\mathbb{E}_t\left[\mathbb{1}_D \cdot G^t\right] = \mathbb{E}_t\left[\mathbb{E}_t[\mathbb{1}_D \cdot G^t|\mathcal{B}_n]\right] = \mathbb{E}_t\left[\mathbb{1}_D \cdot G_n^t\right] = Q(W \in D|T = t).$$

Since this holds for all $D \in \mathcal{E}_n$ and all $n \in \mathbb{N}$ it also holds for all $D \in \mathcal{B}_{\mathcal{W}} = \sigma\left(\bigcup_{n\in\mathbb{N}} \mathcal{E}_n\right)$ and we get:

$$\int \mathbb{1}_D(w) \cdot G(w,t)\, K(W \in dw|T = t) = \mathbb{E}_t\left[\mathbb{1}_D \cdot G^t\right] = Q(W \in D|T = t).$$

Since $Q(W|T = t)$ is a finite measure the set $\{w \in \mathcal{W} \,|\, G(w,t) = \infty\}$ is a $K(W|T = t)$-null set and we can replace G by g under the integral. This shows the claim. □

2.3.4. Transition Probability Spaces

Definition 2.3.23 (Transition probability space)**.** *Consider measurable spaces $\mathcal{T}$ and $\mathcal{W}$ and a Markov kernel:*

$$K(W|T) : \mathcal{T} \dashrightarrow \mathcal{W}, \quad (D,t) \mapsto K(W \in D|T = t).$$

*Then we call the tuple $(\mathcal{W}\times\mathcal{T}, K(W|T))$ a **transition probability space**. It generalizes the notion of probability space, which can be recovered by taking $\mathcal{T} = \ast$.*

Definition 2.3.24 (Conditional random variables)**.** *A measurable map:*

$$X : \mathcal{W} \times \mathcal{T} \to \mathcal{X}$$

*starting from a transition probability space $(\mathcal{W} \times \mathcal{T}, K(W|T))$ is called **conditional random variable**. It generalizes the notion of random variables and can be considered a family of random variables (measurably) parameterized by $t \in \mathcal{T}$. For $t \in \mathcal{T}$ we also define the measurable map:*

$$X_t : \mathcal{W} \to \mathcal{X}, \quad w \mapsto X_t(w) := X(w,t),$$

which can be considered a random variable on the probability space $(\mathcal{W}, K(W|T = t))$.

Example 2.3.25 (Special conditional random variables of importance)**.** *Let $(\mathcal{W} \times \mathcal{T}, K(W|T))$ be a transition probability space. Then we denote by:*

1. *T the **canonical projection** onto $\mathcal{T}$:*

$$T := \mathrm{pr}_{\mathcal{T}} : \mathcal{W} \times \mathcal{T} \to \mathcal{T}, \quad (w,t) \mapsto T(w,t) := t.$$

2. *$\ast$ the **constant conditional random variable**:*

$$\ast : \mathcal{W} \times \mathcal{T} \to \ast, \quad (w,t) \mapsto \ast,$$

where $\ast := \{\ast\}$ is the one-point space.

2.4. Constructing Markov Kernels from Others

2.4.1. Marginal Markov Kernels

Definition 2.4.1 (Marginalizing Markov kernels)**.** *Let*

$$K(X, Y|T) : \mathcal{T} \dashrightarrow \mathcal{X} \times \mathcal{Y}$$

be a Markov kernel in two variables. We can then define the ***marginal Markov kernels*** *as follows:*

$$K(X|T) : \mathcal{T} \dashrightarrow \mathcal{X}, \quad (A, t) \mapsto K(X \in A, Y \in \mathcal{Y}|T = t),$$

and:

$$K(Y|T) : \mathcal{T} \dashrightarrow \mathcal{Y}, \quad (B, t) \mapsto K(X \in \mathcal{X}, Y \in B|T = t).$$

Example 2.4.2 (Marginal Markov kernels of discrete Markov kernels)**.** *Let*

$$K(X, Y|T) : \mathcal{T} \dashrightarrow \mathcal{X} \times \mathcal{Y}$$

be a Markov kernel in two variables on discrete spaces and $k_{X,Y|T}$ its mass function. We can then compute the marginal Markov kernels as follows:

$$k_{X|T}(x|t) = \sum_{y \in \mathcal{Y}} k_{X,Y|T}(x, y|t),$$

and:

$$k_{Y|T}(y|t) = \sum_{x \in \mathcal{X}} k_{X,Y|T}(x, y|t).$$

Note, by abuse of notation, for simplicity, we often omit the indices and write $k(x|t)$ and $k(y|t)$ instead and distinguish these two functions just by the use of the argument symbols x and y.

2.4.2. Product of Markov Kernels

Definition 2.4.3 (Product of Markov kernels)**.** *Consider two Markov kernels:*

$$Q(Z|Y, W, T) : \mathcal{Y} \times \mathcal{W} \times \mathcal{T} \dashrightarrow \mathcal{Z}, \qquad K(W, U|T, X) : \mathcal{T} \times \mathcal{X} \dashrightarrow \mathcal{W} \times \mathcal{U}.$$

Then we define the ***product Markov kernel****:*

$$Q(Z|Y, W, T) \otimes K(W, U|T, X) : \mathcal{Y} \times \mathcal{T} \times \mathcal{X} \dashrightarrow \mathcal{Z} \times \mathcal{W} \times \mathcal{U},$$

using measurable sets $E \subseteq \mathcal{Z} \times \mathcal{W} \times \mathcal{U}$ via: $(E, (y, t, x)) \mapsto$

$$\int \int \mathbb{1}_E(z, w, u)\, Q(Z \in dz|Y = y, W = w, T = t)\, K((W, U) \in d(w, u)|T = t, X = x),$$

where the inner integration is over $z \in \mathcal{Z}$ and the outer integration over $(w, u) \in \mathcal{W} \times \mathcal{U}$.[6]

[6]The integration ordering actually does not matter, which follows from Fubini's theorem, Theorem 2.4.7.

Remark 2.4.4. *Note that in the notation of the product of Markov kernels we use the suggestive symbols, e.g. W in $K(W, U|T, X)$ and W in $Q(Z|Y, W, T)$ to indicate which variables will be "coupled" in the product. A more precise notation could indicate this, e.g. by indices on the product symbol like $\otimes_{(W_1,W_2)}$ or similar. However, our shorthand notation should not lead to much ambiguity during this course. The rule of thumb is that the output of a kernel at the r.h.s. of the product is coupled to a matching input of the kernel on the l.h.s. of the product.*

Example 2.4.5 (Product of discrete Markov kernels)**.** *Let $Q(Z|Y, W, T)$ and $K(W|T, X)$ be two Markov kernels on finite spaces. Let $P(Z, W|Y, T, X) := Q(Z|Y, W, T) \otimes K(W|T, X)$ be the product of Markov kernels and p, q, k the corresponding mass functions. Then we have:*

$$p(z_i, w_k|y_s, x_l, t_j) = q(z_i|y_s, w_k, t_j) \cdot k(w_k|t_j, x_l),$$

which is just the product of mass functions. For the corresponding stochastic tensors $\tilde{P}$, $\tilde{Q}$, $\tilde{K}$ we get that:

$$\tilde{P} = \tilde{Q} \odot_{W,T} \tilde{K}$$

is the entry-wise product/Hadamard product of tensors (reflecting the above formula, i.e. indices for w_k, t_j are the same in q and k).

Exercise 2.4.6. *Show that the product of Markov kernels is associative. Under which conditions can we commute Markov kernels in products? For this you can use Fubini's theorem. See also the comments below.*

Theorem 2.4.7 (Fubini's Theorem, [Kle20] Thm. 14.19)**.** *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two (σ-)finite measure spaces and $f : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ a $(\mathcal{B}_\mathcal{X} \otimes \mathcal{B}_\mathcal{Y})$-$\mathcal{B}_{[0,\infty]}$-measurable map. Then we have the equalities:*

$$\int \left(\int f(x, y)\, \mu(dx) \right) \nu(dy) = \int \left(\int f(x, y)\, \nu(dy) \right) \mu(dx) = \int f(x, y)\, d(\mu \otimes \nu)(x, y).$$

Remark 2.4.8 (Conventions about integration order)**.** *For $D \in \mathcal{B}_\mathcal{X} \otimes \mathcal{B}_\mathcal{Y}$ Fubini's theorem says:*

$$\int \left(\int \mathbb{1}_D(x, y)\, \mu(dx) \right) \nu(dy) = \int \left(\int \mathbb{1}_D(x, y)\, \nu(dy) \right) \mu(dx) = \int \mathbb{1}_D(x, y)\, d(\mu \otimes \nu)(x, y).$$

The integral notation hides the fact that $\mu \otimes \nu$ and $\nu \otimes \mu$ can only be identified as measures if we also swap the order of the spaces $\mathcal{X}$ and $\mathcal{Y}$. $\mu \otimes \nu$ lives on $\mathcal{X} \times \mathcal{Y}$ and $\nu \otimes \mu$ lives on $\mathcal{Y} \times \mathcal{X}$. In more precise terms, we would have:

$$(\mu \otimes \nu)(D) = (\nu \otimes \mu)(D^s),$$

where $D^s := \{(y, x) \in \mathcal{Y} \times \mathcal{X} \mid (x, y) \in D\} \in \mathcal{B}_\mathcal{Y} \otimes \mathcal{B}_\mathcal{X}$. That said, we will always make this swap implicitly, and just write:

$$Q(Z|Y, T) \otimes K(U|T, X) = K(U|T, X) \otimes Q(Z|Y, T),$$

if there is no variable with the same symbol that occurs on the left of the conditioning bar in one Markov kernel and on the right of the conditioning bar in the other Markov kernel. This is justified by Fubini's theorem and our implicit swap convention. This will not lead to much ambiguity, when interpreted as measures under the integral, as one can match the variables Z, U, etc., to their corresponding arguments z, u, etc., in our suggestive notations. In a similar sense we also identify:

$$\begin{aligned}K(W|T=t,X=x) &= K(W|X=x,T=t),\\ Q(X\in A,Y\in B|T) &= Q(Y\in B,X\in A|T).\end{aligned}$$

2.4.3. Composition of Markov Kernels

Definition 2.4.9 (Composition of Markov kernels)**.** *Consider two Markov kernels:*

$$Q(Z|Y,W,T):\mathcal{Y}\times\mathcal{W}\times\mathcal{T}\dashrightarrow\mathcal{Z},\qquad K(W,U|T,X):\mathcal{T}\times\mathcal{X}\dashrightarrow\mathcal{W}\times\mathcal{U}.$$

Then we define their ***composition****:*

$$Q(Z|Y,W,T)\circ K(W,U|T,X):\mathcal{Y}\times\mathcal{T}\times\mathcal{X}\dashrightarrow\mathcal{Z},$$

using measurable sets $C\subseteq\mathcal{Z}$ via:

$$(C,(y,t,x))\quad\mapsto\quad\int Q(Z\in C|Y=y,W=w,T=t)\,K(W\in dw|T=t,X=x).$$

Note that we implicitly marginalized U out, i.e. in the composition we integrate over all variables (here: W and U) from the right hand Markov kernel. As a notation we will also write:

$$\begin{aligned}&Q(Z\in C|Y=y,W,T=t)\circ K(W,U|T=t,X=x)\\ :=\ &\left(Q(Z|Y,W,T)\circ K(W,U|T,X)\right)(C|(y,t,x)).\end{aligned}$$

Remark 2.4.10. *It is clear from the definitions 2.4.9, 2.4.3 and 2.4.1 that the composition:*

$$Q(Z|Y,W,T)\circ K(W,U|T,X)$$

is the Z-marginal of the product:

$$Q(Z|Y,W,T)\otimes K(W,U|T,X).$$

Furthermore, while the operation $\otimes$ leaves all the variables of the second Markov kernel, here W and U, "intact", the operation $\circ$ marginalizes them all out. One could also think of an intermediate operation that specifies which variables are marginalized out and which stays, e.g. using a symbol $\circ_{(W,W)}$ to inticate marginalization of input W (of $Q(Z|Y,W,T)$) over output W (from $K(W,U|T,X)$). We will not further investigate this and will only use $\otimes$ and $\circ$ as described.

Remark 2.4.11 (Composition of deterministic Markov kernels)**.** *Consider measurable maps:*

$$X : \mathcal{T} \to \mathcal{X}, \qquad Z : \mathcal{X} \to \mathcal{Z},$$

and their composition $Z \circ X$. Then the composition of the corresponding Markov kernels satisfies:

$$\delta(Z \circ X|T) = \delta(Z|X) \circ \delta(X|T),$$

where $\delta(Z \in C|X = x) := \mathbb{1}_C(Z(x))$ and $\delta(X \in A|T = t) := \mathbb{1}_A(X(t))$.
So the composition of Markov kernels extends the composition of functions.

Proof.

$$\begin{aligned}
\delta(Z \in C|X) \circ \delta(X|T = t) &= \big(\delta(Z|X) \circ \delta(X|T)\big)(C|t) \\
&= \int \delta(Z \in C|X = x)\, \delta(X \in dx|T = t) \\
&= \int \mathbb{1}_{Z^{-1}(C)}(x)\, \delta(X \in dx|T = t) \\
&= \delta(X \in Z^{-1}(C)|T = t) \\
&= \mathbb{1}_{X^{-1}(Z^{-1}(C))}(t) \\
&= \mathbb{1}_C(Z(X(t))) \\
&= \delta(Z(X) \in C|T = t) \\
&= \delta(Z \circ X \in C|T = t) \\
&= \delta(Z \circ X|T)(C|t).
\end{aligned}$$

□

Example 2.4.12 (Composition of discrete Markov kernels)**.** *Assume that all the spaces in definition 2.4.9 are discrete/finite and let $P(Z|T) := Q(Z|W) \circ K(W|T)$ be the composition of Markov kernels. Let p, q, k denote the corresponding mass functions. Then we get:*

$$p(z_i|t_j) = \sum_k q(z_i|w_k) \cdot k(w_k|t_j).$$

If $\tilde{P}$, $\tilde{Q}$, $\tilde{K}$ are the corresponding stochastic matrices then we have that:

$$\tilde{P} = \tilde{Q}\, \tilde{K},$$

is just the usual matrix product. So in this case the composition of Markov kernels corresponds to matrix multiplication.

2.4.4. Push-Forward of Markov Kernels

Definition 2.4.13 (Push-forward Markov kernel)**.** *Let $(\mathcal{W} \times \mathcal{T}, K(W|T))$ be a transition probability space and:*

$$X : \mathcal{W} \times \mathcal{T} \to \mathcal{X}$$

be a conditional random variable. Then we define the ***push-forward Markov kernel*** $K(X|T)$ *of* $K(W|T)$ *w.r.t.* X *with symbols:*

$$K(X|T) =: X_* K(W|T) =: K(X(W,T)|T),$$

via:

$$K(X|T) : \mathcal{T} \dashrightarrow \mathcal{X}, \quad (A,t) \mapsto K(X \in A|T=t) := K(W \in X_t^{-1}(A)|T=t),$$

where, again:

$$X_t^{-1}(A) = X^{-1}(A)_t := \{w \in \mathcal{W} \,|\, X(w,t) \in A\}.$$

Remark 2.4.14. *We can also write push-forwards as compositions:*

$$K(X|T) = \delta(X|W,T) \circ K(W|T),$$

where we define:

$$\delta(X \in A|W=w, T=t) := \mathbb{1}_A(X(w,t)) = \mathbb{1}_{X^{-1}(A)}(w,t).$$

Remark 2.4.15. *For any Markov kernel*

$$K(W|T) : \mathcal{T} \dashrightarrow \mathcal{W}$$

one can always extend it to include $T = \mathrm{pr}_{\mathcal{T}}$*:*

$$K(W,T|T) : \mathcal{T} \dashrightarrow \mathcal{W} \times \mathcal{T}, \quad (E,t) \mapsto K((W,T) \in E|T=t) = K(W \in E_t|T=t),$$

where $E_t = \{w \in \mathcal{W} \,|\, (w,t) \in E\}$*. Using Definition 2.4.3, we can also write this as:*

$$K(W,T|T) = K(W|T) \otimes \delta(T|T),$$

where $\delta(T \in D|T=t) := \mathbb{1}_D(t)$ *for measurable* $D \subseteq \mathcal{T}$ *and* $t \in \mathcal{T}$*.*

2.4.5. Conditional Markov Kernels

Definition/Theorem 2.4.16 (Disintegration of Markov kernels)**.** *Let* $\mathcal{X}$*,* $\mathcal{Y}$*,* $\mathcal{Z}$ *be measurable spaces where* $\mathcal{X}$ *and* $\mathcal{Y}$ *are standard measurable spaces. Let*

$$K(X,Y|Z) : \mathcal{Z} \dashrightarrow \mathcal{X} \times \mathcal{Y}$$

be a Markov kernel and $K(Y|Z)$ *its marginal Markov kernel given by* $K(Y \in B|Z) = K(X \in \mathcal{X}, Y \in B|Z)$*. Then there exists a Markov kernel (called* ***conditional Markov kernel****):*

$$K(X|Y,Z) : \mathcal{Y} \times \mathcal{Z} \dashrightarrow \mathcal{X}$$

such that:

$$K(X,Y|Z) = K(X|Y,Z) \otimes K(Y|Z).$$

Furthermore, $K(X|Y,Z)$ is essentially unique in the following sense: If $Q(X|Y,Z)$ is another Markov kernel then we have:

$$K(X,Y|Z) = Q(X|Y,Z) \otimes K(Y|Z),$$

if and only if the measurable subset N of $\mathcal{Y} \times \mathcal{Z}$ defined via:

$$N := \{(y,z) \in \mathcal{Y} \times \mathcal{Z} \,|\, \exists A \in \mathcal{B}_{\mathcal{X}} : Q(X \in A|Y=y, Z=z) \neq K(X \in A|Y=y, Z=z)\}$$

is a $K(Y|Z)$-null set in $\mathcal{Y} \times \mathcal{Z}$.

Remark 2.4.17. *If one further assumes certain continuity conditions for the conditional Markov kernel $K(X|Y,Z)$ and that the marginal $K(Y|Z)$ is strictly positive then the conditional Markov kernel can be fully identified, not just up to such $K(Y|Z)$-null sets. This is formalized in Lemma 2.4.23.*

Example 2.4.18 (Conditional Markov kernel for discrete Markov kernels)**.** *Consider a Markov kernel $K(X,Y|Z)$ where all spaces are discrete and let k be the corresponding mass function. Then the marginal mass functions are given by:*

$$k(y|z) = \sum_{x \in \mathcal{X}} k(x,y|z), \qquad k(x|z) = \sum_{y \in \mathcal{Y}} k(x,y|z).$$

A conditional Markov kernel conditioned on Y can then be defined via the mass function:

$$k(x|y,z) := \begin{cases} \frac{k(x,y|z)}{k(y|z)} & \text{if} \quad k(y|z) > 0, \\ k(x|z) & \text{if} \quad k(y|z) = 0.^{7} \end{cases}$$

With this setting we then have for all (!) values x, y, z:

$$k(x,y|z) = k(x|y,z) \cdot k(y|z).$$

Corollary 2.4.19 (Conditional probability distributions)**.** *Let X and Y be random variables on domain $(\mathcal{W}, P(W))$ with standard measurable spaces $\mathcal{X}$, $\mathcal{Y}$, resp., as codomains. Then there always exist* ***conditional probability distributions*** *$P(X|Y)$ and $P(Y|X)$ that are Markov kernels satisfying:*[8]

$$P(X,Y) = P(X|Y) \otimes P(Y), \qquad P(X,Y) = P(Y|X) \otimes P(X).$$

Furthermore, these conditional probability distributions are essentially unique.

---

[7] Any value assignment for this spot is somewhat arbitrary as it almost surely does not occur. Typically this entry is defined to be 0. This is convenient but also problematic, as this would not normalize when summing over $x \in \mathcal{X}$. A proper alternative is to set it to be $k(x|z)$ in this case.

[8] In the literature a conditional probability distribution that is also a Markov kernel would be called a *regular* version of a conditional probability distribution. Since in this lecture we will not encounter other versions we will just call this version here *conditional probability distribution.*

Proofs - Disintegration of Markov Kernels In this subsection we will give a proof for the existence and essential uniqueness of conditional Markov kernels. Another source for similar results can be found in [Kal17].

Remark 2.4.20 (Existence of conditional Markov kernels). *If $K(X, Y|Z)$ is a Markov kernel then we want $K(X|Y, Z)$ such that:*

$$K(X, Y|Z) = K(X|Y, Z) \otimes K(Y|Z)$$

holds. The heuristic here is to make use of Doob-Radon-Nikodym derivatives, see Theorem 2.3.17, for each $A \in \mathcal{B}_{\mathcal{X}}$:

$$K(X \in A|Y = y, Z = z) = \frac{K(X \in A, Y \in dy|Z = z)}{K(Y \in dy|Z = z)}(y).$$

The problem is that they are only unique up to $K(Y|Z)$-null sets and might not be coordinated in such a way that $K(X \in A|Y = y, Z = z)$ becomes a probability measure in A for every (y, z). To ensure this we will take extra steps: We will first take the Doob-Radon-Nikodym derivative $K(X \le x|Y = y, Z = z)$ for rational points $x \in \mathbb{Q}$ and then for general $x \in \mathbb{R}$ put:

$$K(X \le x|Y = y, Z = z) = \inf_{m \in \mathbb{N}} K(X \le \lceil x \rceil_m |Y = y, Z = z),$$

where $\lceil x \rceil_m := \frac{\lfloor mx+1 \rfloor}{m} \in \mathbb{Q}$ for $m \in \mathbb{N}$. This approach will work for $K(Y|Z)$-almost-all (y, z). On the remaining points (y, z) we can then make a somewhat arbitrary choice, e.g. we can put:

$$K(X \le x|Y = y, Z = z) := K(X \le x|Z = z).$$

This will turn $K(X \le x|Y = y, Z = z)$ into a valid cumulative distribution function in x for all (y, z), which then corresponds to a proper probability measure. One then checks that this $K(X|Y, Z)$ is a desired conditional Markov kernel.

Theorem 2.4.21 (Existence of conditional Markov kernels). *Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be measurable spaces where $\mathcal{X}$ is a standard measurable space and $\mathcal{B}_{\mathcal{Y}}$ is countably generated (e.g. $\mathcal{Y}$ is also a standard measurable space). Let*

$$K(X, Y|Z) : \mathcal{Z} \dashrightarrow \mathcal{X} \times \mathcal{Y},$$

be a Markov kernel in two variables. Then a conditional Markov kernel conditioned on Y given Z:

$$K(X|Y, Z) : \mathcal{Y} \times \mathcal{Z} \dashrightarrow \mathcal{X},$$

exists.

Proof. Since $\mathcal{X}$ is standard we can without loss of generality assume that $\mathcal{X} = [0, 1]$. For fixed $A \in \mathcal{B}_{\mathcal{X}}$ we have a finite transition measure $K(X \in A, Y|Z)$ from $\mathcal{Z}$ to $\mathcal{Y}$, which is absolute continuous w.r.t. the marginal $K(Y|Z)$, because of the inequality:

$$0 \le K(X \in A, Y \in B|Z = z) \le K(X \in \mathcal{X}, Y \in B|Z = z) = K(Y \in B|Z = z).$$

Since also $\mathcal{B}_\mathcal{Y}$ is countably generated, by Doob-Radon-Nikodym, see Theorem 2.3.17, we get a Doob-Radon-Nikodym derivative, i.e. a (jointly) measurable map:

$$g_A : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}_{\geq 0},$$

such that for all $z \in \mathcal{Z}$ and $B \in \mathcal{B}_\mathcal{Y}$:

$$K(X \in A, Y \in B|Z = z) = \int \mathbb{1}_B(y) \cdot g_A(y|z)\, K(Y \in dy|Z = z).$$

For $x \in \mathcal{X}$ we will define:

$$G(x|y, z) := g_{[0,x]}(y, z).$$

As a next step we want to modify $G(x|y, z)$ such that it becomes a cumulative distribution function in x, i.e. it corresponds to a probability distribution on $\mathcal{X}$. For this define $\mathcal{X}_\mathbb{Q} := \mathcal{X} \cap \mathbb{Q}$, which is countable and dense in $\mathcal{X}$. First note that:

$$S := \{(y, z) \in \mathcal{Y} \times \mathcal{Z} \,|\, G(1|y, z) \neq 1\}$$

is a measurable $K(Y|Z)$-null set. Then, for every pair $x_1 < x_2$ in $\mathcal{X}_\mathbb{Q}$ consider:

$$E_{(x_1,x_2)} := \{(y, z) \,|\, G(x_1|y, z) > G(x_2|y, z)\} \in \mathcal{B}_\mathcal{Y} \otimes \mathcal{B}_\mathcal{Z}.$$

Since we have the equations:

$$\begin{array}{cl} & \int \mathbb{1}_{E_{(x_1,x_2),z}}(y) \cdot G(x_1|y, z)\, K(Y \in dy|Z = z) \\ = & K(X \leq x_1, Y \in E_{(x_1,x_2),z}|Z = z) \\ \overset{x_1<x_2}{\leq} & K(X \leq x_2, Y \in E_{(x_1,x_2),z}|Z = z) \\ = & \int \mathbb{1}_{E_{(x_1,x_2),z}}(y) \cdot G(x_2|y, z)\, K(Y \in dy|Z = z) \\ \overset{G(x_2|y,z)<G(x_1|y,z)}{\leq} & \int \mathbb{1}_{E_{(x_1,x_2),z}}(y) \cdot G(x_1|y, z)\, K(Y \in dy|Z = z) \end{array}$$

we necessarily have $K(Y \in E_{(x_1,x_2),z}|Z = z) = 0$ for every $z \in \mathcal{Z}$.
Then $E := S \cup \bigcup_{x_1<x_2\in\mathcal{X}_\mathbb{Q}} E_{(x_1,x_2)}$ is also a $K(Y|Z)$-null set in $\mathcal{B}_\mathcal{Y} \otimes \mathcal{B}_\mathcal{Z}$.
Now for $x \in \mathcal{X}_\mathbb{Q}$ we can define:

$$D_x := \{(y, z) \,|\, G(x|y, z) < \inf_{n\in\mathbb{N}} G(x + 1/n|y, z)\} \in \mathcal{B}_\mathcal{Y} \otimes \mathcal{B}_\mathcal{Z}.$$

By the dominated convergence theorem (see [Kle20] Cor. 6.26) we get:

$$\begin{array}{cl} & \int \mathbb{1}_{D_{x,z}}(y) \cdot G(x|y, z)\, K(Y \in dy|Z = z) \\ \overset{D_x}{\leq} & \int \mathbb{1}_{D_{x,z}}(y) \cdot \inf_{n\in\mathbb{N}} G(x + \frac{1}{n}|y, z)\, K(Y \in dy|Z = z) \\ = & \inf_{n\in\mathbb{N}} \int \mathbb{1}_{D_{x,z}}(y) \cdot G(x + \frac{1}{n}|y, z)\, K(Y \in dy|Z = z) \\ = & \inf_{n\in\mathbb{N}} K(X \leq x + \frac{1}{n}, Y \in D_{x,z}|Z = z) \\ = & K(X \leq x, Y \in D_{x,z}|Z = z) \\ = & \int \mathbb{1}_{D_{x,z}}(y) \cdot G(x|y, z)\, K(Y \in dy|Z = z). \end{array}$$

So equality must hold, which then implies that:

$$\int \mathbb{1}_{D_{x,z}}(y) \cdot \underbrace{\left(\inf_{n\in\mathbb{N}} G(x+\frac{1}{n}|y,z) - G(x|y,z) \right)}_{>0 \text{ for } y\in D_{x,z}} K(Y\in dy|Z=z) = 0.$$

This shows that $K(Y \in D_{x,z}|Z = z) = 0$ for all $z \in \mathcal{Z}$. So $D := E \cup \bigcup_{x\in\mathcal{X}_\mathbb{Q}} D_x$ is again a $K(Y|Z)$-null set in $\mathcal{B}_\mathcal{Y} \otimes \mathcal{B}_\mathcal{Z}$.
So far, we got that G, when restricted to $\mathcal{X}_\mathbb{Q} \times D^c$, is jointly measurable in (y, z) for fixed x and monotone non-decreasing and continuous from above in x for fixed (y, z) with $G(1|y, z) = 1$. We now aim to extend G to $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$.
For $x \in \mathcal{X} = [0, 1]$ and $m \in \mathbb{N}$ put $\lceil x\rceil_m := \min(1, \lfloor mx+1\rfloor/m)$. Then $\lceil x\rceil_m \in [0, 1]\cap\mathbb{Q} = \mathcal{X}_\mathbb{Q}$. The map $x \mapsto \lceil x\rceil_m$ is measurable and for $x \in [0, 1)$ we have:

$$x < \lceil x\rceil_m \leq x + \frac{1}{m}.$$

So $\lceil 1\rceil_m = 1$ and $\lceil x\rceil_m \in \mathcal{X}_\mathbb{Q}$ converges to $x \in \mathcal{X}$, $x \neq 1$, from above for $m \to \infty$.
We then define for all $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$:

$$F(x|y,z) := \inf_{m\in\mathbb{N}} \{G(\lceil x\rceil_m|y,z)\} \cdot \mathbb{1}_{D^c}(y,z) + K(X\leq x|Z=z) \cdot \mathbb{1}_D(y,z).$$

It is clear that F is again jointly measurable in (y, z) for fixed x and agrees with G on $\mathcal{X}_\mathbb{Q} \times D^c$ by construction. As a monotone approximation from above it is clearly continuous from above, monotone non-decreasing and satifies $F(1|y, z) = 1$ for all (y, z). So for fixed (y, z) now $F(\cdot|y, z)$ corresponds to a probability distribution $K(X|Y = y, Z = z)$ on $\mathcal{B}_\mathcal{X}$, uniquely given by the defining relations on sets $[0, x]$:

$$F(x|y,z) =: K(X\leq x|Y=y,Z=z),$$

for all $x \in \mathcal{X}$.
Now define $\mathcal{D} \subseteq \mathcal{B}_\mathcal{X}$ as the set of all $A \in \mathcal{B}_\mathcal{X}$ that satisfy:

1. the map $(y, z) \mapsto K(X \in A|Y = y, Z = z)$ is $(\mathcal{B}_\mathcal{Y} \otimes \mathcal{B}_\mathcal{Z})$-$\mathcal{B}_\mathbb{R}$-measurable, and:
2. for all $z \in \mathcal{Z}$ and $B \in \mathcal{B}_\mathcal{Y}$ the following equation holds:

$$K(X\in A, Y\in B|Z=z) = \int \mathbb{1}_B(y) \cdot K(X\in A|Y=y,Z=z)\, K(Y\in dy|Z=z).$$

Since $K(X, Y \in B|Z = z)$ and $K(X|Y = y, Z = z)$ are measures in X the system $\mathcal{D}$ is closed under countable disjoint unions. One can also check that $\mathcal{D}$ is closed under complements and contains $\mathcal{X} = [0, 1]$. So $\mathcal{D}$ is a Dynkin system. We already know that for $x \in \mathcal{X}_\mathbb{Q}$ the map $(y, z) \mapsto K(X \leq x|Y = y, Z = z) = F(x|y, z)$ is measurable. Since for $x \in \mathcal{X}_\mathbb{Q}$ and every $B \in \mathcal{B}_\mathcal{Y}$, $z \in \mathcal{Z}$, we have:

$$\mathbb{1}_B(y) \cdot K(X\leq x|Y=y,Z=z) = \mathbb{1}_B(y) \cdot G(x|y,z)$$

up to the $K(Y|Z=z)$-null set D_z we already get for those $x \in \mathcal{X}_{\mathbb{Q}}$:

$$K(X \leq x, Y \in B|Z=z) = \int \mathbb{1}_B(y) \cdot K(X \leq x|Y=y, Z=z)\, K(Y \in dy|Z=z).$$

This shows that $\mathcal{E} := \{[0,x] \,|\, x \in \mathcal{X}_{\mathbb{Q}}\} \subseteq \mathcal{D}$. Since $\mathcal{E}$ is closed under finite intersections Dynkin's lemma (see [Kle20] Thm. 1.19) implies:

$$\mathcal{B}_{\mathcal{X}} = \sigma(\mathcal{E}) \subseteq \mathcal{D}.$$

This shows that the two conditions hold for all $A \in \mathcal{B}_{\mathcal{X}}$ and thus that $K(X|Y,Z)$ is the desired conditional Markov kernel. □

Lemma 2.4.22 (Essential uniqueness)**.** *If we have Markov kernels:*

$$P(X|Y,Z),\, Q(X|Y,Z) : \mathcal{Y} \times \mathcal{Z} \dashrightarrow \mathcal{X},$$

and

$$K(Y|Z) : \mathcal{Z} \dashrightarrow \mathcal{Y}$$

with any measurable spaces $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ such that:

$$P(X|Y,Z) \otimes K(Y|Z) = Q(X|Y,Z) \otimes K(Y|Z),$$

then for every $A \in \mathcal{B}_{\mathcal{X}}$ the set:

$$N_A := \{(y,z) \in \mathcal{Y} \times \mathcal{Z} \,|\, P(X \in A|Y=y, Z=z) \neq Q(X \in A|Y=y, Z=z)\}$$

is a measurable $K(Y|Z)$-null set.
If, furthermore, $\mathcal{X}$ is countably generated, e.g. a standard measurable space, then also $N := \bigcup_{A \in \mathcal{B}_{\mathcal{X}}} N_A$ is a measurable $K(Y|Z)$-null set.

Proof. For fixed $A \in \mathcal{B}_{\mathcal{X}}$ both $P(X \in A|Y,Z)$ and $Q(X \in A|Y,Z)$ can be considered a Doob-Radon-Nikodym derivative of the same finite transition measure $M_A(Y|Z)$ given by:

$$\begin{aligned} M_A(Y \in B|Z=z) &:= \int \mathbb{1}_B(y) \cdot P(X \in A|Y,Z)\, K(Y \in dy|Z=z) \\ &= \left(P(X \in A|Y,Z) \otimes K(Y|Z)\right)(B|z) \\ &= \left(Q(X \in A|Y,Z) \otimes K(Y|Z)\right)(B|z) \\ &\int \mathbb{1}_B(y) \cdot Q(X \in A|Y,Z)\, K(Y \in dy|Z=z). \end{aligned}$$

The uniqueness statement then follows from that of Doob-Radon-Nikodym derivatives, see Lemma 2.3.21. If now $\mathcal{B}_{\mathcal{X}}$ is countably generated then $\mathcal{B}_{\mathcal{X}} = \sigma(\mathcal{A})$ with a countable set $\mathcal{A}$ that is closed under finite intersections, e.g. $\mathcal{B}_{[0,1]} = \sigma\left(\{[0,c] \,|\, c \in [0,1] \cap \mathbb{Q}\}\right)$.

One then puts $M := \bigcup_{A \in \mathcal{A}} N_A$, which is, as countable union of $K(Y|Z)$-null sets, a $K(Y|Z)$-null set. Then one can define:

$$\mathcal{D} := \{A \in \mathcal{B}_{\mathcal{X}} \,|\, \forall (y,z) \in M^{\mathsf{c}} : \, P(X \in A|Y = y, Z = z) = Q(X \in A|Y = y, Z = z)\}.$$

One easily sees that $\mathcal{D}$ is closed under complements, countable disjoint unions and contains $\mathcal{X}$. This shows that $\mathcal{D}$ is a Dynkin system (aka λ-system). Furthermore, we have: $\mathcal{A} \subseteq \mathcal{D}$ and that $\mathcal{A}$ is closed under finite intersections. By Dynkin's lemma we get that:

$$\mathcal{B}_{\mathcal{X}} = \sigma(\mathcal{A}) \subseteq \mathcal{D}.$$

This shows that $N = \bigcup_{A \in \mathcal{B}_{\mathcal{X}}} N_A \subseteq M$, thus $N = M$ which is a measurable $K(Y|Z)$-null set. □

We now want to prove that we can recover from the ambiguity of the null sets for conditional Markov kernel under continuity assumptions and strictly positive marginals.

Lemma 2.4.23 (Uniqueness for continuous conditional Markov kernels and strictly positive marginals)**.** *Let $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ be Polish spaces endowed with their Borel-σ-algebra and:*

$$P(X|Y,Z),\, Q(X|Y,Z) : \mathcal{Y} \times \mathcal{Z} \to \mathcal{P}(\mathcal{X}),$$

two ***continuous*** *Markov kernels, where $\mathcal{P}(\mathcal{X})$ carries any Hausdorff topology $\mathcal{T}_{\mathcal{X}} \subseteq \mathcal{B}_{\mathcal{P}(\mathcal{X})}$, e.g. the weak*-topology. Let*

$$K(Y|Z) : \mathcal{Z} \dashrightarrow \mathcal{Y}$$

be a Markov kernel that is ***strictly positive*** *(on non-empty open subsets of $\mathcal{Y}$). If we have the equality of Markov kernels $\mathcal{Z} \dashrightarrow \mathcal{X} \times \mathcal{Y}$:*

$$P(X|Y,Z) \otimes K(Y|Z) = Q(X|Y,Z) \otimes K(Y|Z),$$

then we already have the equality of Markov kernels:

$$P(X|Y,Z) = Q(X|Y,Z).$$

Proof. Consider the set:

$$\Delta_{\mathcal{P}(\mathcal{X})} := \{(P,P) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \,|\, P \in \mathcal{P}(\mathcal{X})\},$$

which is a closed subset of $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ because $\mathcal{P}(\mathcal{X})$ is Hausdorff. Then the set:

$$\begin{aligned} N &:= \{(y,z) \in \mathcal{Y} \times \mathcal{Z} \,|\, P(X|Y=y,Z=z) \neq Q(X|Y=y,Z=z)\} \\ &= \left\{(y,z) \in \mathcal{Y} \times \mathcal{Z} \,\middle|\, (P(X|Y=y,Z=z), Q(X|Y=y,Z=z)) \notin \Delta_{\mathcal{P}(\mathcal{X})}\right\} \\ &= (P(X|Y,Z), Q(X|Y,Z))^{-1}(\Delta^{\mathsf{c}}_{\mathcal{P}(\mathcal{X})}), \end{aligned}$$

is an open subset of $\mathcal{Y} \times \mathcal{Z}$ as both Markov kernels are continuous. By the essential uniqueness from Lemma 2.4.22 we know that for all $z \in \mathcal{Z}$ we have:

$$K(Y \in N^z|Z = z) = 0.$$

The fact that the section N^z is open in $\mathcal{Y}$ and that $K(Y|Z)$ is strictly positive implies that either $K(Y \in N^z|Z = z) > 0$ or that $N^z = \emptyset$. Since the former was ruled out by the essential uniqueness we get $N^z = \emptyset$ for all $z \in \mathcal{Z}$ and thus $N = \emptyset$. This shows the claim:

$$P(X|Y, Z) = Q(X|Y, Z).$$

□

2.5. Conditional Independence

2.5.1. Independence for Random Variables

Motivation 2.5.1. *If we throw two dice with outcome values X and Y, resp., then knowing the value of Y does not give us any information about the value of X, and vice versa. We say that X and Y are independent from each other. We will formalize this intuition for all random variables in the following.*

Definition 2.5.2 (Independence of two random variables)**.** *Let $(\mathcal{W}, P(W))$ be a probability space and $X : \mathcal{W} \to \mathcal{X}$ and $Y : \mathcal{W} \to \mathcal{Y}$ be two random variables. We say that X and Y are* ***independent*** *if the following equation holds:*

$$P(X, Y) = P(X) \otimes P(Y),$$

where $P(X, Y)$ is the joint and $P(X)$ and $P(Y)$ are the corresponding marginal distributions. In symbols we would write this as:

$$X \underset{P(W)}{\perp\!\!\!\perp} Y.$$

Lemma 2.5.3. *Let $(\mathcal{W}, P(W))$ be a probability space, $\mathcal{X}$ and $\mathcal{Y}$ standard measurable spaces and $X : \mathcal{W} \to \mathcal{X}$ and $Y : \mathcal{W} \to \mathcal{Y}$ be two random variables. Then the following statements are equivalent:*

1. $X \underset{P(W)}{\perp\!\!\!\perp} Y$.
2. $P(X, Y) = P(X) \otimes P(Y)$.
3. *There exists a probability distribution $Q(X)$ such that:*

$$P(X, Y) = Q(X) \otimes P(Y).$$

4. *$P(X|Y) = P(X)$ holds $P(Y)$-almost-surely, where $P(X|Y)$ is a version of a conditional probability distribution from Cor. 2.4.19.*
5. *For all $A \in \mathcal{B}_{\mathcal{X}}$ we have:*

$$\mathbb{E}[\mathbb{1}_A(X)|Y] = \mathbb{E}[\mathbb{1}_A(X)] \qquad P(W)\text{-a.s.}$$

6. *For all $A \in \mathcal{B}_\mathcal{X}$ and $B \in \mathcal{B}_\mathcal{Y}$ we have:*

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B).$$

Proof. Exercise. $\square$

Exercise 2.5.4. *Reformulate the statements in Lemma 2.5.3 for the case we either have mass functions (discrete case) or densities w.r.t. a product measure, e.g. the Lebesgue measure (absolute continuous case).*

We can generalize the notion of independence to arbitrary families of random variables:

Definition 2.5.5 (Mutual independence of families of random variables)**.** *Let $(\mathcal{W}, P(W))$ be a probability space and I an (index) set. For $i \in I$ let $X_i : \mathcal{W} \to \mathcal{X}_i$ be a random variable. We say that $(X_i)_{i \in I}$ is **(mutually/jointly) independent** if for all two disjoint subsets $J_1 \dot\cup J_2 \subseteq I$ we have the independence:*

$$(X_{j_1})_{j_1 \in J_1} \underset{P(W)}{\perp\!\!\!\perp} (X_{j_2})_{j_2 \in J_2}.$$

Exercise 2.5.6 (Mutual independence for finite tuples of random variables)**.** *A finite tuple of random variables $(X_1, \dots, X_n)$ is mutually independent if and only if:*

$$P(X_1, \dots, X_n) = P(X_1) \otimes \cdots \otimes P(X_n).$$

Exercise 2.5.7. *Let $(\mathcal{W}, P(W))$ be a probability space and I an arbitrary index set. For $i \in I$ let $X_i : \mathcal{W} \to \mathcal{X}_i$ be a random variable. The following statements are equivalent:*

1. *$(X_i)_{i \in I}$ is (mutually/jointly) independent.*

2. *For every finite disjoint subsets $J_1, J_2 \subseteq I$ we have the independence:*

$$(X_{j_1})_{j_1 \in J_1} \underset{P(W)}{\perp\!\!\!\perp} (X_{j_2})_{j_2 \in J_2}.$$

3. *For every finite subset $J \subseteq I$ and $i \in I \setminus J$ we have:*

$$X_i \underset{P(W)}{\perp\!\!\!\perp} (X_j)_{j \in J}.$$

4. *For every finite subset $J \subseteq I$ we have:*

$$P\left((X_j)_{j \in J}\right) = \bigotimes_{j \in J} P(X_j).$$

5. *We have the equality:*

$$P\left((X_i)_{i \in I}\right) = \bigotimes_{i \in I} P(X_i),$$

where $\bigotimes_{i \in I} P(X_i)$ is the product measure on $\mathcal{X} = \prod_{i \in I} \mathcal{X}_i$, which is determined by the corresponding products on its finite marginals via the extension theorem of Ionescu-Tulcea, see [IT49, Lam87] and theorem A.10.2.

2.5.2. Conditional Independence for Random Variables

Motivation 2.5.8. *1. Consider two independent coin flips with outcome variables X and Y, resp., with values in $\{0,1\}$, and $Z := X + Y \in \{0,1,2\}$. If the value of Z is known, say $Z = 1$, then revealing the value of Y, say $Y = 0$, provides us with all the information to fully determine the value of X, here $X = 0$. This is despite the fact that X and Y were assumed to be independent. This means that conditioning on a third variable Z can destroy independence. In this case, we say that X and Y are dependent conditioned on Z. Summarized in symbols we have:*

$$X \underset{P(W)}{\perp\!\!\!\perp} Y, \qquad \textit{but:} \qquad X \underset{P(W)}{\not\!\perp\!\!\!\perp} Y \,|\, Z.$$

*2. Now consider three (mutually) independent coin flips X, W, U with values in $\{0,1\}$. Let $Z := X + W$ and $Y := Z + U = X + W + U$. If we knew the value of Y, say $Y = 0$, then we would have information about the values of X as well, here $X = 0$. This shows that X and Y can not be independent random variables. If, in contrast, we would first reveal the value of Z, say $Z = 1$, then the value of X might be restricted by the value of Z, but also revealing Y would not give us any **additional** information about the value of X. The reason is that $Y = Z + U$ and U is independent of X, W and $Z = X + W$. So, even though X and Y are dependent, when conditioned on Z they become independent, as there is no additional information gained about each others value, when revealing the other. Summarized in symbols we have:*

$$X \underset{P(W)}{\not\!\perp\!\!\!\perp} Y, \qquad \textit{but:} \qquad X \underset{P(W)}{\perp\!\!\!\perp} Y \,|\, Z.$$

We now want to formalize conditional independence for random variables.

Remark 2.5.9 (Conditional independence)**.** *In contrast to (unconditional) independence, see Definition 2.5.2, possible definitions of conditional independence come with many more subtleties, due to their interplay with conditional probability distributions or conditional expectations. Such definitions can in general be non-equivalent. However, if we restrict ourselves to standard measurable spaces the subtleties can be resolved and the definitions become equivalent. This is the reason that in the following we will only state conditional independence for standard measurable spaces. We will make a clearer choice later for conditional independence of conditional random variables.*

Definition/Lemma 2.5.10 (Conditional independence for random variables)**.** *Let $(\mathcal{W}, P(W))$ be a probability space and $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ standard measurable spaces, and $X : \mathcal{W} \to \mathcal{X}$ and $Y : \mathcal{W} \to \mathcal{Y}$ and $Z : \mathcal{W} \to \mathcal{Z}$ be three random variables. We then say that X **is independent of** Y **conditioned on** Z if any of the following equivalent conditions holds:*

1. *$P(X,Y|Z) = P(X|Z) \otimes P(Y|Z)$ holds $P(Z)$-a.s., where $P(X,Y|Z)$, $P(X|Z)$, $P(Y|Z)$ are versions of conditional probability distributions from Cor. 2.4.19.*

2. *$P(X|Y,Z) = P(X|Z)$ holds $P(Y,Z)$-a.s., where $P(X|Y,Z)$, $P(X|Z)$ from Cor. 2.4.19.*

3. *There exists a Markov kernel $Q(X|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}$ such that:*

$$P(X,Y,Z) = Q(X|Z) \otimes P(Y,Z).$$

4. *For every $A \in \mathcal{B}_\mathcal{X}$ and $B \in \mathcal{B}_\mathcal{Y}$ we have:*

$$\mathbb{E}\left[\mathbb{1}_A(X) \cdot \mathbb{1}_B(Y)|Z\right] = \mathbb{E}\left[\mathbb{1}_A(X)|Z\right] \cdot \mathbb{E}\left[\mathbb{1}_B(Y)|Z\right] \qquad P(W)\text{-a.s.}$$

5. *For every $A \in \mathcal{B}_\mathcal{X}$ we have:*

$$\mathbb{E}\left[\mathbb{1}_A(X)|Y,Z\right] = \mathbb{E}\left[\mathbb{1}_A(X)|Z\right] \qquad P(W)\text{-a.s.}$$

In those cases, in symbols we write:

$$X \underset{P(W)}{\perp\!\!\!\perp} Y \,|\, Z.$$

Proof. Exercise. □

Exercise 2.5.11. *Restate all the statements in Definition/Lemma 2.5.10 for discrete random variables in terms of mass functions.*

Conditional independence for random variables satisfies the following rules:

Theorem 2.5.12 (Separoid axioms for conditional independence for random variables, see [Daw01])**.** *Let $(\mathcal{W}, P(W))$ be a probability space and X, Y, Z and U random variables taking values in standard measurable spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ and $\mathcal{U}$, respectively. Then we have the following rules:*

1. *Redundancy: If $U = \varphi(X)$ a.s. is a measurable function of X, e.g. $U = X$, then:*

$$U \underset{P(W)}{\perp\!\!\!\perp} Y \,|\, X.$$

2. *Symmetry:*

$$X \underset{P(W)}{\perp\!\!\!\perp} Y \,|\, Z \qquad \Longrightarrow \qquad Y \underset{P(W)}{\perp\!\!\!\perp} X \,|\, Z.$$

3. *Decomposition:*

$$X \underset{P(W)}{\perp\!\!\!\perp} Y, U \,|\, Z \qquad \Longrightarrow \qquad X \underset{P(W)}{\perp\!\!\!\perp} U \,|\, Z.$$

4. *Weak Union:*

$$X \underset{P(W)}{\perp\!\!\!\perp} Y, U \,|\, Z \qquad \Longrightarrow \qquad X \underset{P(W)}{\perp\!\!\!\perp} Y \,|\, U, Z.$$

5. *Contraction:*

$$\left(X \underset{P(W)}{\perp\!\!\!\perp} U \,|\, Z\right) \quad \wedge \quad \left(X \underset{P(W)}{\perp\!\!\!\perp} Y \,|\, U, Z\right) \qquad \Longrightarrow \qquad X \underset{P(W)}{\perp\!\!\!\perp} Y, U \,|\, Z.$$

Proof. Exercise. □

Remark 2.5.13. *The separoid axioms, see Theorem 2.5.12, also hold true for random variables that map into general (non-standard) measurable spaces if one restricts oneself to the definition of conditional independence only involving conditional expectations (rather than conditional probabilities or Markov kernels) in Definition 2.5.10.*

Exercise 2.5.14. *Assume that the random variables X, Y, Z, U have a joint density p w.r.t. some product measure (or a joint mass function) such that $p(y, u|z) > 0$ for all values y, u, z. Show that we then also have the following* ***intersection*** *rule:*

$$\left(X \underset{P(W)}{\perp\!\!\!\perp} U \,|\, Y, Z\right) \quad \wedge \quad \left(X \underset{P(W)}{\perp\!\!\!\perp} Y \,|\, U, Z\right) \qquad \Longrightarrow \qquad X \underset{P(W)}{\perp\!\!\!\perp} Y, U \,|\, Z.$$

2.5.3. Conditional Independence for Conditional Random Variables

Motivation 2.5.15. *Assume that we are given a statistical model $P(X|\Theta)$ and a statistic $S = S(X)$, which is a measurable function of X. Often one wants to find such an S such that the choice of parameter $\Theta = \theta$ has no "influence" on the probability distribution of X when S is provided. Such a statistic is usually called a* ***sufficient*** *statistic of X w.r.t. $P(X|\Theta)$. In symbols we want S such that:*

$$X \perp\!\!\!\perp \Theta \,|\, S.$$

However, the parameter variable Θ here is not a proper random variable as we have no distribution $P(\Theta)$ specified over it. Still such a conditional independence statement makes sense. We thus want to formalize a notion of conditional independence for conditional random variables. We follow the definition of [For21]. Other approaches can be found in [Daw79, Daw80, Daw01, CD17, RERS23, FM20].

Motivation 2.5.16. *Consider a probabilistic program with input variables T, S and output variables X, Y, Z. Whenever the program is given T and S as input, it internally samples $U, E \sim U[0, 1]$ uniformly and independently from a random number generator, then calculates:*

$$X := T + S + U, \qquad\qquad Y := 5 \cdot S + E, \qquad\qquad Z := X \cdot Y,$$

and, finally, outputs X, Y and Z. Even though, the input T and S is provided by the user and is not considered a random variable, we can reason about the fact that "Output Y only depends on the input S and not on the input T." We want to formalize such conditional indpendence mathematically in order to be able to write this as:

$$Y \perp\!\!\!\perp T \,|\, S.$$

Definition 2.5.17 (Conditional independence for conditional random variables)**.** *Let* $(\mathcal{W} \times \mathcal{T}, K(W|T))$ *be a transition probability space with Markov kernel:*

$$K(W|T) : \mathcal{T} \dashrightarrow \mathcal{W}.$$

Consider conditional random variables:

$$X : \mathcal{W} \times \mathcal{T} \to \mathcal{X}, \qquad Y : \mathcal{W} \times \mathcal{T} \to \mathcal{Y}, \qquad Z : \mathcal{W} \times \mathcal{T} \to \mathcal{Z}.$$

We say that X ***is independent of*** Y ***conditioned on*** Z ***w.r.t.*** $K(W|T)$*, in symbols:*

$$X \underset{K(W|T)}{\perp\!\!\!\perp} Y \,|\, Z,$$

if there ***exists*** *a Markov kernel:*

$$Q(X|Z) : \mathcal{Z} \dashrightarrow \mathcal{X},$$

such that:

$$K(X, Y, Z|T) = Q(X|Z) \otimes K(Y, Z|T),$$

where $K(Y, Z|T)$ *is the marginal of* $K(X, Y, Z|T)$*.*[9]
As a special case, we define:

$$X \underset{K(W|T)}{\perp\!\!\!\perp} Y \qquad :\Longleftrightarrow \qquad X \underset{K(W|T)}{\perp\!\!\!\perp} Y \,|\, *.$$

Notation 2.5.18 (Essential uniqueness)**.** *The Markov kernel* $Q(X|Z)$ *appearing in the conditional independence* $X \perp\!\!\!\perp_{K(W|T)} Y \,|\, Z$ *in definition 2.5.17 is then a version of a conditional Markov kernel* $K(X|Y, Z, T)$ *and is thus essentially unique in the sense of 2.4.22. We will use the following suggestive notation for it:*

$$K(X|\cancel{T, Y}, Z) := Q(X|Z),$$

or similarly with crossed variables in different order. So we have in case of $X \perp\!\!\!\perp_{K(W|T)} Y \,|\, Z$*:*

$$K(X, Y, Z|T) = K(X|\cancel{T, Y}, Z) \otimes K(Y, Z|T).$$

Note that $K(X|\cancel{T, Y}, Z)$ *is a version of the conditional Markov kernel* $K(X|Y, Z, T)$ *and does not depend on arguments* y *and* t*.*

Remark 2.5.19 (Conditional independence includes conditional independence from T)**.** *We have the equivalence:*

$$X \underset{K(W|T)}{\perp\!\!\!\perp} Y \,|\, Z \quad \Longleftrightarrow \quad X \underset{K(W|T)}{\perp\!\!\!\perp} T, Y \,|\, Z,$$

where $T : \mathcal{W} \times \mathcal{T} \to \mathcal{T}$*,* $(w, t) \mapsto t$*, is the canonical projection map.*

[9] For the equation $K(X, Y, Z|T) = Q(X|Z) \otimes K(Y, Z|T)$ to hold it is sufficient to check that for all $t \in \mathcal{T}$, $A \in \mathcal{B}_{\mathcal{X}}$, $B \in \mathcal{B}_{\mathcal{Y}}$ and $C \in \mathcal{B}_{\mathcal{Z}}$ we have:

$$K(X \in A, Y \in B, Z \in C|T = t) = \int_C \int_B Q(X \in A|Z = z)\, K(Y \in dy, Z \in dz|T = t).$$

Proof.

$$
\begin{aligned}
& X \underset{K(W|T)}{\perp\!\!\!\perp} Y \mid Z \\
\iff\ & \exists Q(X|Z):\ K(X,Y,Z|T) && = Q(X|Z) \otimes K(Y,Z|T) \\
\iff\ & \exists Q(X|Z):\ K(X,T,Y,Z|T) && = Q(X|Z) \otimes \underbrace{K(Y,Z|T) \otimes \delta(T|T)}_{K(T,Y,Z|T)} \\
\iff\ & X \underset{K(W|T)}{\perp\!\!\!\perp} T, Y \mid Z.
\end{aligned}
$$

The middle implication "$\Longrightarrow$" follows by taking the product with $\delta(T|T)$, and the reverse implication "$\Longleftarrow$" by marginalizing out T, i.e. via $\delta(T \in \mathcal{T}|T) = 1$. $\square$

Remark 2.5.20 (How to find $Q(X|Z)$ and check for conditional independence?)**.** *In case we have the conditional independence:*

$$X \underset{K(W|T)}{\perp\!\!\!\perp} Y \mid Z,$$

we then get by definition:

$$K(X,Y,Z|T) = Q(X|Z) \otimes K(Y,Z|T),$$

for some Markov kernel $Q(X|Z)$. This implies for all $t \in \mathcal{T}$ the equation:

$$K(X,Z|T=t) = Q(X|Z) \otimes K(Z|T=t).$$

This means that $Q(X|Z)$ is a version of the conditional probability distribution $K(X|Z,T=t)$ for all $t \in \mathcal{T}$ at once, and, in addition, it is also functionally not dependent on t. So for fixed $t_0 \in \mathcal{T}$ the conditional $K(X|Z,T=t_0)$ can be changed on a $K(Z|T=t_0)$-null-set such that it agrees with $Q(X|Z)$. So it is reasonable to test out versions of $K(X|Z,T=t_0)$ for $Q(X|Z)$. To summarize, we have the following equivalence between:

1. $X \underset{K(W|T)}{\perp\!\!\!\perp} Y \mid Z$,
2. *There exist $t_0 \in \mathcal{T}$ and a (regular) version of the conditional probability distribution $K(X|Z,T=t_0)$ such that for all $t \in \mathcal{T}$:*

   $$K(X,Y,Z|T=t) = K(X|Z,T=t_0) \otimes K(Y,Z|T=t).$$

   Note that in the last expression the middle term has the fixed t_0 and the outer two terms have varying $t \in \mathcal{T}$.

This equivalence allows us to narrow our search space for $Q(X|Z)$ to such conditional probability distributions.

Example 2.5.21 (Conditional independence for discrete conditional random variables)**.** *Let the situation be like in definition 2.5.17 and assume all spaces to be countable and discrete. Let k be the mass function for $K(X, Y, Z|T)$. Then we have:*

$$X \underset{K(W|T)}{\perp\!\!\!\perp} Y \mid Z,$$

if and only if there is a probability mass function q such that for all values x, y, z, t:

$$k(x, y, z|t) = q(x|z) \cdot k(y, z|t).$$

Note that in this case $q(x|z)$ is a version of $k(x|z, t)$ for all $t \in \mathcal{T}$ at once, but that is also independent of t. We can use this knowledge to find such a proposal $q(x|z)$ as follows.

If there exists a $t_0 \in \mathcal{T}$ such that $k(z|t_0) > 0$ for all $z \in \mathcal{Z}$ then the conditional $k(x|z, t_0)$ is uniquely given and equal to $\frac{k(x,z|t_0)}{k(z|t_0)}$. $k(x|z, t_0)$ would then necessarily agree with $q(x|z)$ in case of the conditional independence. So, if there exists a $t_0 \in \mathcal{T}$ such that $k(z|t_0) > 0$ for all $z \in \mathcal{Z}$ then we get the following equivalence:

$$X \underset{K(W|T)}{\perp\!\!\!\perp} Y \mid Z \qquad \iff \qquad \forall x, y, z, t : \qquad k(x, y, z|t) = k(x|z, t_0) \cdot k(y, z|t).$$

Again, note that in the last expression the middle term has the fixed t_0 and the outer two terms have varying $t \in \mathcal{T}$.

This example can be generalized.

Theorem 2.5.22 (Conditional independence for conditional random variables with density)**.** *Let $\mu_{\mathcal{X}}$, $\mu_{\mathcal{Y}}$ and $\mu_{\mathcal{Z}}$ reference measures on $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, resp., and $\mu := \mu_{\mathcal{X}} \otimes \mu_{\mathcal{Y}} \otimes \mu_{\mathcal{Z}}$ the product measure on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Assume that $K(X, Y, Z|T)$ has a Doob-Radon-Nikodym derivative k w.r.t. μ and let $t_0 \in \mathcal{T}$ be a fixed value. Then we have the implication:*

$$\forall t \in \mathcal{T}\, \forall_\mu\, x, y, z. \qquad k(x, y, z|t) = k(x|z, t_0) \cdot k(y, z|t) \qquad \Longrightarrow \qquad X \underset{K(W|T)}{\perp\!\!\!\perp} Y \mid Z,$$

where $\forall_\mu$ means "for μ-almost-all". If for $\mu_{\mathcal{Z}}$-almost-all $z \in \mathcal{Z}$ we have: $k(z|t_0) > 0$, then also the reverse implication holds, with $k(x|z, t_0) := \frac{k(x,z|t_0)}{k(z|t_0)}$:

$$\forall t \in \mathcal{T}\, \forall_\mu\, x, y, z. \qquad k(x, y, z|t) = k(x|z, t_0) \cdot k(y, z|t) \qquad \Longleftarrow \qquad X \underset{K(W|T)}{\perp\!\!\!\perp} Y \mid Z.$$

Proof. " $\Longrightarrow$ ": This is clear, as the factorization of the densities provides the needed factorization of the corresponding Markov kernels.
"$\Longleftarrow$": By assumption we have a factorization:

$$K(X, Y, Z|T) = Q(X|Z) \otimes K(Y, Z|T),$$

which implies for all $A \in \mathcal{B}_\mathcal{X}$ and $C \in \mathcal{B}_\mathcal{Z}$:

$$\begin{aligned}
&\int_C \int_A k(x,z|t_0)\, \mu_\mathcal{X}(dx)\, \mu_\mathcal{Z}(dz) \\
&= K(X \in A, Z \in C | T = t_0) \\
&= \int_C Q(X \in A | Z = z)\, K(Z \in dz | T = t_0) \\
&= \int_C Q(X \in A | Z = z)\, k(z|t_0)\, \mu_\mathcal{Z}(dz).
\end{aligned}$$

This implies that we have:

$$\forall_{\mu_\mathcal{Z}} z \in \mathcal{Z}. \qquad \int_A k(x,z|t_0)\, \mu_\mathcal{X}(dx) = Q(X \in A | Z = z)\, k(z|t_0),$$

which implies, since $k(z|t_0) > 0$, that $k(x|z,t_0) = \frac{k(x,z|t_0)}{k(z|t_0)}$ is a density of $Q(X|Z)$ up to a $\mu_\mathcal{Z}$-null set N. Since $K(Z|T) \ll \mu_\mathcal{Z}$ this N is also a $K(Z|T)$-null set, and thus $\mathcal{Y} \times N$ a $K(Y,Z|T)$-null set. So for all $t \in \mathcal{T}$, $A \in \mathcal{B}_\mathcal{X}$, $B \in \mathcal{B}_\mathcal{Y}$, $C \in \mathcal{B}_\mathcal{Z}$ we get:

$$\begin{aligned}
&\int_C \int_B \int_A k(x,y,z|t)\, \mu_\mathcal{X}(dx)\, \mu_\mathcal{Y}(dy)\, \mu_\mathcal{Z}(dz) \\
&= K(X \in A, Y \in B, Z \in C | T = t) \\
&= \int_C \int_B Q(X \in A | Z = z)\, K(Y \in dy, Z \in dz | T = t) \\
&= \int_C \int_B \int_A k(x|z,t_0)\, \mu_\mathcal{X}(dx)\, K(Y \in dy, Z \in dz | T = t) \\
&= \int_C \int_B \int_A k(x|z,t_0)\, \mu_\mathcal{X}(dx)\, k(y,z|t)\, \mu_\mathcal{Y}(dy)\, \mu_\mathcal{Z}(dz) \\
&= \int_C \int_B \int_A k(x|z,t_0) \cdot k(y,z|t)\, \mu_\mathcal{X}(dx)\, \mu_\mathcal{Y}(dy)\, \mu_\mathcal{Z}(dz).
\end{aligned}$$

So the corresponding Markov kernels on the lhs and rhs are the same. This implies that the set:

$$M := \{(x,y,z,t) \,|\, k(x,y,z|t) \neq k(x|z,t_0) \cdot k(y,z|t)\}$$

is a μ-null set in $\mathcal{B}_\mathcal{X} \otimes \mathcal{B}_\mathcal{Y} \otimes \mathcal{B}_\mathcal{Z} \otimes \mathcal{B}_\mathcal{T}$. So for all $t \in \mathcal{T}$ and μ-almost-all x, y, z the following equation holds:

$$k(x,y,z|t) = k(x|z,t_0) \cdot k(y,z|t),$$

which implies the claim. □

Remark 2.5.23 (Conditional independence for random variables)**.** *By Definition/Lemma 2.5.10 we recover the notion of conditional independence for random variables X, Y, Z with standard measurable spaces as codomains by taking $\mathcal{T} := \{*\}$, $P(W) := K(W|*)$:*

$$X \underset{P(W)}{\perp\!\!\!\perp} Y \,|\, Z \qquad \iff \qquad \exists Q(X|Z) :\; P(X,Y,Z) = Q(X|Z) \otimes P(Y,Z).$$

Such a $Q(X|Z)$ is then a conditional probability distribution of $P(X,Z)$ conditioned on Z. In suggestive notations:

$$Q(X|Z) =: P(X|\cancel{Y}, Z) = P(X|Z).$$

Lemma 2.5.24 (Conditional independence for deterministic mappings)**.** *Let $F : \mathcal{T} \to \mathcal{F}$ and $H : \mathcal{T} \to \mathcal{H}$ be measurable mappings, with $\mathcal{F}$ standard. We now consider them as (deterministic) conditional random variables on the transition probability space $(\mathcal{W} \times \mathcal{T}, K(W|T))$ via:*

$$\begin{array}{rccc} F: & \mathcal{W} \times \mathcal{T} & \to & \mathcal{F}, \\ & (w,t) & \mapsto & F(t), \\ H: & \mathcal{W} \times \mathcal{T} & \to & \mathcal{H}, \\ & (w,t) & \mapsto & H(t), \end{array}$$

which do not depend on the 'probabilistic part' $\mathcal{W}$ of $K(W|T)$. Let $G : \mathcal{W} \times \mathcal{T} \to \mathcal{G}$ be another conditional random variable.
We write $F \precsim H$ if there exists a measurable map $\varphi : \mathcal{H} \to \mathcal{F}$ such that:

$$F = \varphi \circ H.$$

Then we have the equivalence:[10]

$$F \precsim H \qquad \iff \qquad F \underset{K(W|T)}{\perp\!\!\!\perp} G \,|\, H.$$

So F is a deterministic measurable map of H iff F is independent of G given H. Note that the first part of the statement is independent of G.

Proof. "$\implies$": This direction is rather easy. See the later separoid axioms.
"$\Longleftarrow$": Since F and H are deterministic and only dependent on T we get that:

$$K(F,G,H|T) = \delta(F|T) \otimes \delta(H|T) \otimes K(G|T).$$

By the conditional independence we now have a Markov kernel $Q(F|H)$ such that we have the factorization:

$$K(F,G,H|T) = Q(F|H) \otimes K(G,H|T) = Q(F|H) \otimes \delta(H|T) \otimes K(G|T).$$

Marginalizing out G, H and taking $T = t$ we get from these equations:

$$\delta_{F(t)} = \delta(F|T = t) = Q(F|H(t)),$$

which is a Dirac measure centered at $F(t)$. We can now define the mapping:

$$\varphi : \; H(\mathcal{T}) \to \mathcal{F}, \quad H(t) \mapsto F(t),$$

[10]The full equivalence needs Kuratowski's extension theorem for standard measurable spaces (see [Kec95] 12.2): Any measurable map from a (not necessarily measurable) subset of a measurable space to a standard measurable space extends to a measurable map on the whole space. Alternatively, one could define $F \precsim H$ via existence of measurable $\varphi : H(\mathcal{T}) \to \mathcal{F}$ such that $F = \varphi \circ H$, but this moves problems elsewhere.

which is well-defined, because $h := H(t_1) = H(t_2)$ implies that $Q(F|H = h)$ is a Dirac measure centered at $F(t_1)$ and $F(t_2)$, so $F(t_1) = F(t_2)$. φ is measurable. Indeed, its composition with $\delta : \mathcal{F} \to \mathcal{P}(\mathcal{F})$, $z \mapsto \delta_z$ equals $Q(F|H)$, which is measurable. Since $\mathcal{B}_\mathcal{F} = \delta^* \mathcal{B}_{\mathcal{P}(\mathcal{F})}$, see lemma 2.7.1 2., also φ is measurable. Since $\mathcal{F}$ is a standard measurable space, φ extends to a measurable mapping $\varphi : \mathcal{H} \to \mathcal{F}$ by Kuratowski's extension theorem for standard measurable spaces (see [Kec95] 12.2). Finally, note that we have $F(t) = \varphi(H(t))$ for all $(w, t) \in \mathcal{W} \times \mathcal{T}$, which shows the claim. □

Example 2.5.25 (Conditional independence for deterministic mappings)**.** *If for example, $\mathcal{T} = \mathcal{T}_1 \times \mathcal{T}_2$ and $T_i : \mathcal{W} \times \mathcal{T}_1 \times \mathcal{T}_2 \to \mathcal{T}_i$ the canonical projection onto $\mathcal{T}_i$, then F is a function in two variables (t_1, t_2). We then have:*

$$F \underset{K(W|T)}{\perp\!\!\!\perp} T_1 \,|\, T_2,$$

if and only if F—as a function—is only dependent on the argument t_2 (and not on t_1).

Another example of what conditional independence of conditional random variables can encode is the following.

Remark 2.5.26 (Existence of conditional Markov kernels expressed as conditional independence)**.** *Let X, Y be conditional random variables on transition probability space $(\mathcal{W} \times \mathcal{T}, K(W|T))$. Then we can express the existence of a conditional Markov kernel $K(X|Y, T)$ as the conditional independence:*

$$X \underset{K(W|T)}{\perp\!\!\!\perp} * | Y, T,$$

where $*$ is the constant conditional random variable. Alternatively and equivalently, we could also write:

$$X \underset{K(W|T)}{\perp\!\!\!\perp} T | Y, T.$$

Note that for standard measurable spaces $\mathcal{X}$ and $\mathcal{Y}$ the above statement always holds. In suggestive symbols:

$$K(X|\cancel{T}, Y, T) = K(X|Y, T).$$

Example 2.5.27 (Certain statistics expressed as conditional independence)**.** *Let $P(W|\Theta)$ be a statistical model, considered as a Markov kernel $\mathcal{F} \dashrightarrow \mathcal{W}$. Let X and Y be two conditional random variables w.r.t. $P(W|\Theta)$. A* ***statistic*** *of X is a measurable map $S : \mathcal{X} \to \mathcal{S}$, which we consider as the conditional random variable $S \precsim X$ given via:*

$$S : \mathcal{W} \times \mathcal{F} \to \mathcal{S}, \quad (w, \theta) \mapsto S(X(w, \theta)).$$

1. ***Ancillarity****. S is an* ***ancillary statistic*** *of X w.r.t. Θ if and only if:*

$$S \underset{P(W|\Theta)}{\perp\!\!\!\perp} \Theta.$$

This means that every parameter $\Theta = \theta$ induces the same distribution for S:

$$P(S|\Theta = \theta) = P(S|\cancel{\Theta}).$$

2. ***Sufficiency***. *S is a* ***sufficient statistic*** *of X w.r.t.* Θ *if and only if:*

$$X \underset{P(W|\Theta)}{\perp\!\!\!\perp} \Theta \,|\, S.$$

This means that there is a Markov kernel $P(X|S,\cancel{\Theta})$ *such that:*

$$P(X,S|\Theta) = P(X|S,\cancel{\Theta}) \otimes P(S|\Theta).$$

So X only "interacts" with the parameters Θ *through S.*

3. ***Adequacy***. *S is an* ***adequate statistic*** *of X for Y w.r.t.* Θ *if and only if:*

$$X \underset{P(W|\Theta)}{\perp\!\!\!\perp} \Theta, Y \,|\, S.$$

This means we have a factorization:

$$P(X,Y,S|\Theta) = P(X|\cancel{\Theta,Y},S) \otimes P(Y,S|\Theta),$$

for some Markov kernel $P(X|\cancel{\Theta,Y},S)$*. This means that all information of X about the parameters and labels Y is fully captured already by S.*

Theorem 2.5.28. *Let* $(\mathcal{W} \times \mathcal{T}, K(W|T))$ *be a transition probability space with Markov kernel:*

$$K(W|T) : \mathcal{T} \dashrightarrow \mathcal{W}.$$

Consider conditional random variables X, Y, Z with common domain $\mathcal{W} \times \mathcal{T}$ *and codomains* $\mathcal{X}$, $\mathcal{Y}$ *and* $\mathcal{Z}$*, resp., and* $T : \mathcal{W} \times \mathcal{T} \to \mathcal{T}$ *the canonical projection map. We will write* $P(X|Z) = K(X|\cancel{T},Z)$ *for a fixed version of the Markov kernel appearing in the conditional independence* $X \perp\!\!\!\perp_{K(W|T)} T \,|\, Z$ *(only in case it holds).*
With these notations, the following are equivalent:

1. $X \underset{K(W|T)}{\perp\!\!\!\perp} Y \,|\, Z,$
2. $X \underset{K(W|T)}{\perp\!\!\!\perp} T, Y \,|\, Z,$
3. $X \underset{K(W|T)}{\perp\!\!\!\perp} T \,|\, Z$ *and* $K(X,Y,Z|T) = P(X|Z) \otimes K(Y,Z|T)$.
4. $X \underset{K(W|T)}{\perp\!\!\!\perp} T \,|\, Z$ *and for every* $t \in \mathcal{T}$ *we have:* $X_t \underset{K(W|T=t)}{\perp\!\!\!\perp} Y_t \,|\, Z_t$.

Furthermore, any of those points implies:

$$K(X|\cancel{T,Y},Z) = K(X|\cancel{T},Z) \qquad K(Y,Z|T)\text{-a.s.}.$$

and the following:

5. *For every probability distribution $Q(T) \in \mathcal{P}(\mathcal{T})$ we have the conditional independence*[11]*:*

$$X \underset{K(W|T)\otimes Q(T)}{\perp\!\!\!\perp} T, Y \mid Z.$$

Proof. 3. $\implies$ 1. is clear by definition.
1. $\iff$ 2.: by 2.5.19.
2. $\implies$ 4.,5.: By assumption we have the factorization:

$$K(X, Y, Z, T|T) = K(X|Z) \otimes K(Y, Z, T|T),$$

for some Markov kernel $K(X|Z)$. Via marginalization and multiplication this implies the two equations:

$$\begin{aligned} K(X, Z, T|T) &= K(X|Z) \otimes K(Z, T|T), \\ \underbrace{K(X, Y, Z|T) \otimes Q(T)}_{=:Q(X,Y,Z,T)} &= K(X|Z) \otimes \underbrace{K(Y, Z|T) \otimes Q(T)}_{=Q(Y,Z,T)}, \end{aligned}$$

for every $Q(T) \in \mathcal{P}(\mathcal{T})$. The last equation shows 5.
If we take $Q(T) = \delta_t$ we get:

$$K(X_t, Y_t, Z_t|T = t) = K(X|Z_t) \otimes K(Y_t, Z_t|T = t).$$

Together with the first of the above equations this shows 4.
4. $\implies$ 3.: By $X \perp\!\!\!\perp_{K(W|T)} T \mid Z$ we have:

$$K(X, Z|T) = P(X|Z) \otimes K(Z|T).$$

By the assumption $X_t \perp\!\!\!\perp_{K(W|T=t)} Y_t \mid Z_t$, on the other hand, we have—for each $t \in \mathcal{T}$ individually—a factorization:

$$K(X, Y, Z|T = t) = Q_t(X|Z) \otimes K(Y, Z|T = t),$$

with a Markov kernel Q_t, which might depend on $t \in \mathcal{T}$, where we suppress the indices t on all the variables for readability everywhere. Marginalizing out Y and comparing to the above we then get the two equalities:

$$P(X|Z) \otimes K(Z|T = t) = K(X, Z|T = t) = Q_t(X|Z) \otimes K(Z|T = t).$$

By the essential uniqueness of such a factorization we see that $P(X \in A|Z)$ only differs from $Q_t(X \in A|Z)$ on a $K(Z|T = t)$-null set. Considered as functions of (y, z) (by ignoring y) they are equal up to a $K(Y, Z|T = t)$-null set. This means that we can

[11] Note that this again implies the second part of point 4: $X_t \perp\!\!\!\perp Y_t \mid Z_t$ for every $t \in \mathcal{T}$. So the first part of point 4: $X \perp\!\!\!\perp T \mid Z$, can then be seen as the additional obstruction to obtain the "full" conditional independence: $X \perp\!\!\!\perp Y \mid Z$.

replace $Q_t(X \in A|Z)$ with $P(X \in A|Z)$ for every $A \in \mathcal{B}_\mathcal{X}$ and $t \in \mathcal{T}$. So we get the equation:

$$K(X, Y, Z|T = t) = P(X|Z) \otimes K(Y, Z|T = t),$$

for all $t \in \mathcal{T}$ and thus:

$$K(X, Y, Z|T) = P(X|Z) \otimes K(Y, Z|T).$$

This shows 3.

$\square$

Remark 2.5.29 (Discrete $\mathcal{T}$)**.** *In the setting of theorem 2.5.28, let $\mathcal{X}$, $\mathcal{Z}$ be standard measurable spaces and $\mathcal{T}$ be a countable discrete measurable space with any fixed probability distribution $Q(T)$ that has a strictly positive mass function. Then we get the equivalence:*

$$X \underset{K(W|T)\otimes Q(T)}{\perp\!\!\!\perp} T, Y \,|\, Z \qquad \iff \qquad X \underset{K(W|T)}{\perp\!\!\!\perp} Y \,|\, Z.$$

Proof. The rhs implies the lhs side by theorem 2.5.28. So, now assume the lhs and put:

$$Q(X, Y, Z, T) := K(X, Y, Z|T) \otimes Q(T).$$

Its marginal is then denoted by $Q(X, Z)$. Since $\mathcal{X}$, $\mathcal{Z}$ are standard measurable spaces we get a (regular) conditional probability distribution $Q(X|Z)$, such that $Q(X, Z) = Q(X|Z) \otimes Q(Z)$. By the assumed conditional independence we thus have:

$$\begin{aligned} K(X, Y, Z|T) \otimes Q(T) &= Q(X, Y, Z, T) \\ &= Q(X|Z) \otimes Q(Y, Z, T) \\ &= Q(X|Z) \otimes K(Y, Z|T) \otimes Q(T). \end{aligned}$$

Since $Q(T)$ is strictly positive and conditional Markov kernels are essentially unique we get the sure equality:

$$K(X, Y, Z|T) = Q(X|Z) \otimes K(Y, Z|T),$$

which implies the claim. $\square$

Corollary 2.5.30. *If $\mathcal{X}$, $\mathcal{Z}$ are standard measurable spaces then we have the equivalence:*

$$X \underset{K(W|T)}{\perp\!\!\!\perp} Y \,|\, Z, T \qquad \iff \qquad \forall t \in \mathcal{T}: \quad X_t \underset{K(W|T=t)}{\perp\!\!\!\perp} Y_t \,|\, Z_t.$$

Proof. This directly follows from theorem 2.5.28 4. with (Z, T) in the role of Z and remark 2.5.26 to get the first part of 4. In suggestive symbols:

$$K(X|\cancel{T, Y}, Z, T) = K(X|Z, T) \qquad K(Z|T)\text{-a.s.}$$

$\square$

2.5.4. Example: Linear Gaussian Markov Kernels

Theorem 2.5.31 (Conditional independence for linear Gaussian conditional random variables)**.** *Let $\mathcal{T} = \mathbb{R}$, $\mathcal{X} = \mathbb{R}^{d_X}$, $\mathcal{Y} = \mathbb{R}^{d_Y}$ and $\mathcal{Z} = \mathbb{R}^{d_Z}$. Consider a linear Gaussian Markov kernel $P(X, Y, Z|T)$, which is given by a density of the form:*

$$p(x, y, z|t) = \mathcal{N}\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix} \middle| \begin{bmatrix} \Gamma_X \\ \Gamma_Y \\ \Gamma_Z \end{bmatrix} \cdot t + \begin{bmatrix} \gamma_X \\ \gamma_Y \\ \gamma_Z \end{bmatrix}, \begin{bmatrix} \Sigma_{X,X} & \Sigma_{X,Y} & \Sigma_{X,Z} \\ \Sigma_{Y,X} & \Sigma_{Y,Y} & \Sigma_{Y,Z} \\ \Sigma_{Z,X} & \Sigma_{Z,Y} & \Sigma_{Z,Z} \end{bmatrix} \right).$$

Then we have the following equivalence:

$$X \underset{P(X,Y,Z|T)}{\perp\!\!\!\perp} Y \mid Z \qquad \iff \qquad \Sigma_{X,Y} = \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y} \quad \wedge \quad \Gamma_X = \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}\Gamma_Z.$$

If this is the case then the Markov kernel $Q(X|Z)$ coming from the conditional independence:

$$P(X, Y, Z|T) = Q(X|Z) \otimes P(Y, Z|T),$$

is also a linear Gaussian Markov kernel with density:

$$\begin{aligned} q(x|z) &= \mathcal{N}\left(x \,\middle|\, \mu_{X|Z}(z), \Sigma_{X,X|Z}\right), \\ \mu_{X|Z}(z) &:= \gamma_X + \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}(z - \gamma_Z), \\ \Sigma_{X,X|Z} &:= \Sigma_{X,X} - \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,X}, \end{aligned}$$

which coincides with the usual marginal conditional for $t = 0$, i.e.:

$$Q(X|Z = z) = P(X|Z = z, T = 0).$$

So, we also get the equivalence:

$$X \underset{P(X,Y,Z|T)}{\perp\!\!\!\perp} Y \mid Z \qquad \iff \qquad P(X, Y, Z|T) = P(X|Z, T = 0) \otimes P(Y, Z|T).$$

Proof. First note that in general the conditional $P(X|Y, Z, T)$ is also a linear Gaussian Markov kernel and of the form:

$$p(x|y, z, t) = \mathcal{N}\left(x \,\middle|\, \mu_{X|Y,Z,T}(y, z, t), \Sigma_{X|Y,Z,T}\right),$$

with the following abbreviation for the covariance matrix:

$$\Sigma_{X|Y,Z,T} := \Sigma_{X,X} - \begin{bmatrix} \Sigma_{X,Y} & \Sigma_{X,Z} \end{bmatrix} \begin{bmatrix} \Sigma_{Y,Y} & \Sigma_{Y,Z} \\ \Sigma_{Z,Y} & \Sigma_{Z,Z} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{Y,X} \\ \Sigma_{Z,X} \end{bmatrix},$$

and the following abbreviation for the mean:

$$\begin{aligned} &\mu_{X|Y,Z,T}(y, z, t) \\ &:= (\Gamma_X \cdot t + \gamma_X) + \begin{bmatrix} \Sigma_{X,Y} & \Sigma_{X,Z} \end{bmatrix} \begin{bmatrix} \Sigma_{Y,Y} & \Sigma_{Y,Z} \\ \Sigma_{Z,Y} & \Sigma_{Z,Z} \end{bmatrix}^{-1} \left(\begin{bmatrix} y \\ z \end{bmatrix} - \begin{bmatrix} \Gamma_Y \cdot t + \gamma_Y \\ \Gamma_Z \cdot t + \gamma_Z \end{bmatrix} \right) \\ &= \begin{bmatrix} \Sigma_{X,Y} & \Sigma_{X,Z} \end{bmatrix} \begin{bmatrix} \Sigma_{Y,Y} & \Sigma_{Y,Z} \\ \Sigma_{Z,Y} & \Sigma_{Z,Z} \end{bmatrix}^{-1} \begin{bmatrix} y \\ z \end{bmatrix} \\ &\qquad + (\Gamma_X \cdot t + \gamma_X) - \begin{bmatrix} \Sigma_{X,Y} & \Sigma_{X,Z} \end{bmatrix} \begin{bmatrix} \Sigma_{Y,Y} & \Sigma_{Y,Z} \\ \Sigma_{Z,Y} & \Sigma_{Z,Z} \end{bmatrix}^{-1} \begin{bmatrix} \Gamma_Y \cdot t + \gamma_Y \\ \Gamma_Z \cdot t + \gamma_Z \end{bmatrix}. \end{aligned}$$

We now want to investigate under which conditions we get the conditional independence:

$$X \underset{P(X,Y,Z|T)}{\perp\!\!\!\perp} Y \mid Z.$$

Note that in this case the conditional independence is equivalent to the statement that the conditional density $p(x|y,z,t)$ is not dependent on the arguments y and t.

Let us first investigate the first term involving y:

$$\begin{bmatrix} \Sigma_{X,Y} & \Sigma_{X,Z} \end{bmatrix} \begin{bmatrix} \Sigma_{Y,Y} & \Sigma_{Y,Z} \\ \Sigma_{Z,Y} & \Sigma_{Z,Z} \end{bmatrix}^{-1} \begin{bmatrix} y \\ z \end{bmatrix}.$$

Note that we can use the following formula for the (2×2)-block inverse:

$$\begin{bmatrix} \Sigma_{Y,Y} & \Sigma_{Y,Z} \\ \Sigma_{Z,Y} & \Sigma_{Z,Z} \end{bmatrix}^{-1} =$$
$$\begin{bmatrix} \left(\Sigma_{Y,Y} - \Sigma_{Y,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\right)^{-1} & -\left(\Sigma_{Y,Y} - \Sigma_{Y,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\right)^{-1}\Sigma_{Y,Z}\Sigma_{Z,Z}^{-1} \\ -\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\left(\Sigma_{Y,Y} - \Sigma_{Y,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\right)^{-1} & \Sigma_{Z,Z}^{-1} + \Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\left(\Sigma_{Y,Y} - \Sigma_{Y,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\right)^{-1}\Sigma_{Y,Z}\Sigma_{Z,Z}^{-1} \end{bmatrix}$$

This leads us to require that:

$$\begin{bmatrix} \Sigma_{X,Y} & \Sigma_{X,Z} \end{bmatrix} \begin{bmatrix} \left(\Sigma_{Y,Y} - \Sigma_{Y,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\right)^{-1} \\ -\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\left(\Sigma_{Y,Y} - \Sigma_{Y,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\right)^{-1} \end{bmatrix} = 0,$$

which is equivalent to:

$$\left(\Sigma_{X,Y} - \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\right)\left(\Sigma_{Y,Y} - \Sigma_{Y,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\right)^{-1} = 0,$$

which can be further simplified, by multiplying with the inverse of the inverse, to:

$$\Sigma_{X,Y} - \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y} = 0.$$

We also need that the mean of the conditional is not dependent on t, which leads to the following condition coming from the second term:

$$\begin{aligned} 0 &= \Gamma_X - \begin{bmatrix} \Sigma_{X,Y} & \Sigma_{X,Z} \end{bmatrix} \begin{bmatrix} \Sigma_{Y,Y} & \Sigma_{Y,Z} \\ \Sigma_{Z,Y} & \Sigma_{Z,Z} \end{bmatrix}^{-1} \begin{bmatrix} \Gamma_Y \\ \Gamma_Z \end{bmatrix} \\ &= \Gamma_X - \left(\Sigma_{X,Y} - \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\right)\left(\Sigma_{Y,Y} - \Sigma_{Y,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\right)^{-1}\Gamma_Y \\ &\quad - \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}\Gamma_Z + \left(\Sigma_{X,Y} - \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\right)\left(\Sigma_{Y,Y} - \Sigma_{Y,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y}\right)^{-1}\Sigma_{Y,Z}\Sigma_{Z,Z}^{-1}\Gamma_Z \\ &= \Gamma_X - \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}\Gamma_Z, \end{aligned}$$

where we made repeated use of the condition: $\Sigma_{X,Y} - \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y} = 0$.

This leads us to the following equivalence for linear Gaussian Markov kernels:

$$X \underset{P(X,Y,Z|T)}{\perp\!\!\!\perp} Y \mid Z \qquad \iff \qquad \Sigma_{X,Y} = \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,Y} \quad \wedge \quad \Gamma_X = \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}\Gamma_Z.$$

If this is the case then the Markov kernel $Q(X|Z)$ coming from the conditional independence:

$$P(X,Y,Z|T) = Q(X|Z) \otimes P(Y,Z|T),$$

is also a linear Gaussian Markov kernel with density:

$$\begin{aligned} q(x|z) &= \mathcal{N}\left(x \,\middle|\, \mu_{X|Z}(z), \Sigma_{X,X|Z}\right), \\ \mu_{X|Z}(z) &:= \gamma_X + \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}(z-\gamma_Z), \\ \Sigma_{X,X|Z} &:= \Sigma_{X,X} - \Sigma_{X,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,X}, \end{aligned}$$

which is the usual marginal conditional for $t = 0$. □

2.6. Separoid Axioms for Conditional Independence

The following asymmetric separiod axioms for conditional independence are a generalization of the symmetric separoid axioms due to A.P. Dawid [Daw01] and the similar graphoid axioms due to J. Pearl and A. Paz [PP85].

Definition/Theorem 2.6.1 ((Asymmetric) separoid axioms for conditional independence)**.** *Let $(\mathcal{W} \times \mathcal{T}, K(W|T))$ be a transition probability space with Markov kernel:*

$$K(W|T) : \mathcal{T} \dashrightarrow \mathcal{W}.$$

Consider conditional random variables X, Y, Z, U with common domain $\mathcal{W} \times \mathcal{T}$ and standard measurable spaces $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, $\mathcal{U}$, resp., as codomains. Let $T : \mathcal{W} \times \mathcal{T} \to \mathcal{T}$ be the canonical projection and $*$ the constant conditional random variable.
We write $U \precsim X$ if there exists a measurable function $G : \mathcal{X} \to \mathcal{U}$ such that $U = G \circ X$. Then the ternary relation $\perp\!\!\!\perp = \perp\!\!\!\perp_{K(W|T)}$ satisfies the following rules:

a) Extended Left Redundancy:

$U \precsim X \implies U \perp\!\!\!\perp Y \,|\, X.$

b) T-Restricted Right Redundancy:[12]

$X \perp\!\!\!\perp * | Z, T$ *always holds.*

c) T-Inverted Right Decomposition:

$X \perp\!\!\!\perp Y \,|\, Z \implies X \perp\!\!\!\perp T, Y \,|\, Z.$

d) Left Decomposition:

$X, U \perp\!\!\!\perp Y \,|\, Z \implies U \perp\!\!\!\perp Y \,|\, Z.$

e) Right Decomposition:

[12] T-Restricted Right Redundancy, Left Weak Union and Symmetry need the existence of conditional Markov kernels. That is the reason we assumed standard measurable spaces.

$$X \perp\!\!\!\perp Y, U \mid Z \implies X \perp\!\!\!\perp U \mid Z.$$

f) Left Weak Union:[12]

$$X, U \perp\!\!\!\perp Y \mid Z \implies X \perp\!\!\!\perp Y \mid U, Z.$$

g) Right Weak Union:

$$X \perp\!\!\!\perp Y, U \mid Z \implies X \perp\!\!\!\perp Y \mid U, Z.$$

h) Left Contraction:

$$(X \perp\!\!\!\perp Y \mid U, Z) \wedge (U \perp\!\!\!\perp Y \mid Z) \implies X, U \perp\!\!\!\perp Y \mid Z.$$

i) Right Contraction:

$$(X \perp\!\!\!\perp Y \mid U, Z) \wedge (X \perp\!\!\!\perp U \mid Z) \implies X \perp\!\!\!\perp Y, U \mid Z.$$

j) Right Cross Contraction:

$$(X \perp\!\!\!\perp Y \mid U, Z) \wedge (U \perp\!\!\!\perp X \mid Z) \implies X \perp\!\!\!\perp Y, U \mid Z.$$

k) Flipped Left Cross Contraction:

$$(X \perp\!\!\!\perp Y \mid U, Z) \wedge (Y \perp\!\!\!\perp U \mid Z) \implies Y \perp\!\!\!\perp X, U \mid Z.$$

In particular, we have the equivalences:

$$(X \perp\!\!\!\perp Y, U \mid Z) \iff (X \perp\!\!\!\perp Y \mid U, Z) \quad \wedge \quad (X \perp\!\!\!\perp U \mid Z),$$

$$(X, U \perp\!\!\!\perp Y \mid Z) \iff (X \perp\!\!\!\perp Y \mid U, Z) \quad \wedge \quad (U \perp\!\!\!\perp Y \mid Z).$$

We also get:

l) T-Restricted Symmetry:[12]

$$X \perp\!\!\!\perp Y \mid Z, T \implies Y \perp\!\!\!\perp X \mid Z, T.$$

In the special case of $\mathcal{T} = * = \{*\}$, the one-point space, (i.e. in the case of probability distributions and random variables mapping to standard measurable spaces) we thus have (unrestricted) Symmetry.

Proofs - Separoid Axioms for Conditional Independence

In the following let $(\mathcal{W} \times \mathcal{T}, K(W|T))$ be a transition probability space with Markov kernel:

$$K(W|T) : \mathcal{T} \dashrightarrow \mathcal{W},$$

and conditional random variables X, Y, Z, U with common domain $\mathcal{W} \times \mathcal{T}$ and measurable spaces $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, $\mathcal{U}$, resp., as codomains. We indicate when we need to assume standard measurable spaces.
We will use $T : \mathcal{W} \times \mathcal{T} \to \mathcal{T}$ to denote the canonical projection and $*$ to denote the constant conditional random variable.
Recall that we write $U \precsim X$ if there exists a measurable function $\varphi : \mathcal{X} \to \mathcal{U}$ such that $U = \varphi \circ X$.
Recall that for proving:

$$X \underset{K(W|T)}{\perp\!\!\!\perp} Y \mid Z,$$

we need to find/construct a Markov kernel $Q(X|Z)$ such that:

$$K(X, Y, Z|T) = Q(X|Z) \otimes K(Y, Z|T),$$

which is equivalent to:
For all $t \in \mathcal{T}$ and all measurable $A \subseteq \mathcal{X}$, $B \subseteq \mathcal{Y}$, $C \subseteq \mathcal{Z}$ we have the equation:

$$K(X \in A, Y \in B, Z \in C|T = t) = \int_C \int_B Q(X \in A|Z = z)\, K(Y \in dy, Z \in dz|T = t).$$

We abbreviate $\perp\!\!\!\perp := \perp\!\!\!\perp_{K(W|T)}$ in the following.

Lemma 2.6.2 (Extended Left Redundancy)**.**

$$U \precsim X \implies U \perp\!\!\!\perp Y \mid X.$$

Proof. If $U = \varphi(X)$ put $Q(U \in D|X = x) := \delta_\varphi(U \in D|X = x) := \mathbb{1}_D(\varphi(x))$ for $D \subseteq \mathcal{U}$. Then we get:

$$\begin{aligned}
&\int_C \int_B Q(U \in D|X = x)\, K(Y \in dy, X \in dx|T = t) \\
&= \int_C \int_B \delta_\varphi(U \in D|X = x)\, K(Y \in dy, X \in dx|T = t) \\
&= \int_C \int_B \mathbb{1}_{\varphi^{-1}(D)}(x)\, K(Y \in dy, X \in dx|T = t) \\
&= K(Y \in B, X \in C \cap \varphi^{-1}(D)|T = t) \\
&= K(Y \in B, X \in C, \varphi(X) \in D|T = t) \\
&= K(U \in D, Y \in B, X \in C|T = t).
\end{aligned}$$

This shows the claim. In suggestive symbols:

$$K(U|\cancel{T, Y}, X) = \delta_\varphi(U|X).$$

$\square$

Lemma 2.6.3 (T-Restricted Right Redundancy)**.** *Let $\mathcal{X}$ and $\mathcal{Z}$ be standard measurable spaces. Then:*

$$X \perp\!\!\!\perp * \,|\, Z, T \qquad \textit{holds.}$$

Proof. Because $\mathcal{X}$ and $\mathcal{Z}$ are standard measurable spaces we have a factorization:

$$K(X, *, Z, T|T) = K(X|Z, T) \otimes K(*, Z, T|T).$$

with the conditional Markov kernel $K(X|Z, T)$ of $K(X, Z|T)$ (via theorem 2.4.16). This already shows the claim. In suggestive symbols:

$$K(X|\cancel{*, T}, Z, T) = K(X|Z, T).$$

$\square$

Lemma 2.6.4 (T-Inverted Right Decomposition)**.**

$$X \perp\!\!\!\perp Y \,|\, Z \implies X \perp\!\!\!\perp T, Y \,|\, Z.$$

Proof. By assumption we have:

$$K(X, Y, Z|T) = Q(X|Z) \otimes K(Y, Z|T).$$

Multiplying both sides with $\delta(T|T)$ we get:

$$K(X, Y, Z, T|T) = Q(X|Z) \otimes K(T, Y, Z|T).$$

This shows the claim using the same $Q(X|Z)$. In suggestive symbols:

$$K(X|\cancel{T, T, Y}, Z) = K(X|\cancel{T, Y}, Z).$$

$\square$

Lemma 2.6.5 (Left Decomposition)**.**

$$X, U \perp\!\!\!\perp Y \,|\, Z \implies U \perp\!\!\!\perp Y \,|\, Z.$$

Proof. Let $Q(X, U|Z)$ be given from the left conditional independence. Then we have:

$$K(X, U, Y, Z|T) = Q(X, U|Z) \otimes K(Y, Z|T).$$

Marginalizing out X gives:

$$K(U, Y, Z|T) = Q(U|Z) \otimes K(Y, Z|T).$$

This shows the claim. In suggestive symbols:

$$K(U|\cancel{T, Y}, Z) = K(X \in \mathcal{X}, U|\cancel{T, Y}, Z).$$

$\square$

Lemma 2.6.6 (Right Decomposition)**.**

$$X \perp\!\!\!\perp Y, U \,|\, Z \implies X \perp\!\!\!\perp U \,|\, Z.$$

Proof. Let $Q(X|Z)$ be given from the left conditional independence. We then have:

$$K(X, U, Y, Z|T) = Q(X|Z) \otimes K(Y, U, Z|T).$$

Marginalizing out Y gives:

$$K(X, U, Z|T) = Q(X|Z) \otimes K(U, Z|T).$$

This shows the claim. In suggestive symbols:

$$K(X|\cancel{T, U}, Z) = K(X|\cancel{T, Y, U}, Z).$$

$\square$

Lemma 2.6.7 (Left Weak Union)**.** *Let $\mathcal{X}$ and $\mathcal{U}$ be standard measurable spaces. Then:*

$$X, U \perp\!\!\!\perp Y \,|\, Z \implies X \perp\!\!\!\perp Y \,|\, U, Z.$$

Proof. By assumption we have:

$$K(X, U, Y, Z|T) = Q(X, U|Z) \otimes K(Y, Z|T),$$

for some Markov kernel $Q(X, U|Z)$. If we marginalize out X we get:

$$K(U, Y, Z|T) = Q(U|Z) \otimes K(Y, Z|T).$$

Because $\mathcal{X}$ and $\mathcal{U}$ are standard measurable spaces we have a factorization:

$$Q(X, U|Z) = Q(X|U, Z) \otimes Q(U|Z).$$

with the conditional Markov kernel $Q(X|U, Z)$ (via theorem 2.4.16).
Putting these equations together we get:

$$\begin{aligned} K(X, U, Y, Z|T) &= Q(X, U|Z) \otimes K(Y, Z|T) \\ &= Q(X|U, Z) \otimes Q(U|Z) \otimes K(Y, Z|T) \\ &= Q(X|U, Z) \otimes K(U, Y, Z|T). \end{aligned}$$

In suggestive symbols, this means that: $K(X|\cancel{T, Y}, U, Z)$ is the conditional of $K(X, U|\cancel{T, Y}, Z)$.
$\square$

Lemma 2.6.8 (Right Weak Union)**.**

$$X \perp\!\!\!\perp Y, U \,|\, Z \implies X \perp\!\!\!\perp Y \,|\, U, Z.$$

Proof. We have the factorization:

$$K(X,Y,U,Z|T) = Q(X|Z) \otimes K(Y,U,Z|T),$$

with some Markov kernel $Q(X|Z)$. If we view $Q(X|Z)$ as a function in (u,z) via:

$$(u,z) \mapsto Q(X|Z=z),$$

by just ignoring the argument u then the claim follows from the same factorization above.
In suggestive symbols:

$$K(X|\cancel{T,Y},U,Z) = K(X|\cancel{T,Y,U},Z).$$

□

Lemma 2.6.9 (Left Contraction)**.**

$$(X \perp\!\!\!\perp Y \mid U,Z) \wedge (U \perp\!\!\!\perp Y \mid Z) \implies X,U \perp\!\!\!\perp Y \mid Z.$$

Proof. By assumption we have the two factorizations:

$$\begin{aligned} K(X,Y,U,Z|T) &= Q(X|U,Z) \otimes K(Y,U,Z|T), \\ K(Y,U,Z|T) &= P(U|Z) \otimes K(Y,Z|T), \end{aligned}$$

with some Markov kernels $Q(X|U,Z)$, $P(U|Z)$. Putting these equations together using $Q(X|U,Z) \otimes P(U|Z)$ we get:

$$K(X,Y,U,Z|T) = (Q(X|U,Z) \otimes P(U|Z)) \otimes K(Y,Z|T).$$

In suggestive symbols:

$$K(X,U|\cancel{T,Y},Z) = K(X|\cancel{T,Y},U,Z) \otimes K(U|\cancel{T,Y},Z).$$

□

Lemma 2.6.10 (Right Contraction)**.**

$$(X \perp\!\!\!\perp Y \mid U,Z) \wedge (X \perp\!\!\!\perp U \mid Z) \implies X \perp\!\!\!\perp Y,U \mid Z.$$

Proof. By assumption we have the two factorizations:

$$\begin{aligned} K(X,Y,U,Z|T) &= Q(X|U,Z) \otimes K(Y,U,Z|T), \\ K(X,U,Z|T) &= P(X|Z) \otimes K(U,Z|T), \end{aligned}$$

with some Markov kernels $Q(X|U,Z)$, $P(X|Z)$.
Marginalizing out Y we get the equalities:

$$\begin{aligned} K(X,U,Z|T) &= Q(X|U,Z) \otimes K(U,Z|T), \\ K(X,U,Z|T) &= P(X|Z) \otimes K(U,Z|T). \end{aligned}$$

By the essential uniqueness (see lemma 2.4.22) of such factorization we get that for every $A \in \mathcal{B}_{\mathcal{X}}$:

$$Q(X \in A|U, Z) = P(X \in A|Z) \qquad K(U, Z|T)\text{-a.s.}$$

The same equation then holds also $K(Y, U, Z|T)$-a.s. (by ignoring argument y). Plugging that back into the first equation gives:

$$K(X, Y, U, Z|T) = P(X|Z) \otimes K(Y, U, Z|T).$$

In suggestive symbols:

$$K(X|\cancel{T, Y, U}, Z) = K(X|\cancel{T, Y}, U, Z) = K(X|\cancel{T, U}, Z) \qquad \text{a.s.}$$

$\square$

Lemma 2.6.11 (Right Cross Contraction)**.**

$$(X \perp\!\!\!\perp Y \mid U, Z) \wedge (U \perp\!\!\!\perp X \mid Z) \implies X \perp\!\!\!\perp Y, U \mid Z.$$

Proof. By assumption we have the two factorizations:

$$K(X, Y, U, Z|T) = Q(X|U, Z) \otimes K(Y, U, Z|T), \tag{3}$$
$$K(X, U, Z|T) = P(U|Z) \otimes K(X, Z|T), \tag{4}$$

with some Markov kernels $Q(X|U, Z)$, $P(U|Z)$.
We then define the Markov kernel:

$$R(X, U|Z) := Q(X|U, Z) \otimes P(U|Z). \tag{5}$$

We will now show that its marginal:

$$R(X|Z) = Q(X|U, Z) \circ P(U|Z). \tag{6}$$

will satisfy the claim.
If we marginalize out Y from equation 3 we get:

$$K(X, U, Z|T) = Q(X|U, Z) \otimes K(U, Z|T). \tag{7}$$

Equating equations 4 and 7 gives:

$$P(U|Z) \otimes K(X, Z|T) = K(X, U, Z|T) = Q(X|U, Z) \otimes K(U, Z|T). \tag{8}$$

Marginalizing out X in equation 8 on both sides gives:

$$K(U, Z|T) = P(U|Z) \otimes K(Z|T). \tag{9}$$

If we now plug equation 9 into 7 then we get:

$$K(X, U, Z|T) = Q(X|U, Z) \otimes P(U|Z) \otimes K(Z|T) \tag{10}$$
$$\stackrel{5}{=} R(X, U|Z) \otimes K(Z|T). \tag{11}$$

If we marginalize out U in equation 11 and use equation 6 we arrive at:

$$K(X, Z|T) = R(X|Z) \otimes K(Z|T). \tag{12}$$

We now get:

$$\begin{aligned} Q(X|U, Z) \otimes K(U, Z|T) &\overset{7}{=} K(X, U, Z|T) && (13)\\ &\overset{4}{=} P(U|Z) \otimes K(X, Z|T) && (14)\\ &\overset{12}{=} P(U|Z) \otimes R(X|Z) \otimes K(Z|T) && (15)\\ &= R(X|Z) \otimes P(U|Z) \otimes K(Z|T) && (16)\\ &\overset{9}{=} R(X|Z) \otimes K(U, Z|T). && (17) \end{aligned}$$

By the essential uniqueness (see lemma 2.4.22) of such a factorization we get that for every $A \in \mathcal{B}_{\mathcal{X}}$:

$$Q(X \in A|U, Z) = R(X \in A|Z) \qquad K(U, Z|T)\text{-a.s.} \tag{18}$$

The same equation then holds also $K(Y, U, Z|T)$-a.s. (by ignoring the non-occurring argument y). Plugging equation 18 back into the equation 3 we get:

$$\begin{aligned} K(X, Y, U, Z|T) &= Q(X|U, Z) \otimes K(Y, U, Z|T), && (19)\\ &= R(X|Z) \otimes K(Y, U, Z|T). && (20) \end{aligned}$$

This shows the claim.
In suggestive symbols:

$$K(X|\cancel{T, Y, U}, Z) = K(X|\cancel{T, Y}, U, Z) \circ K(U|\cancel{T, X}, Z).$$

$\square$

Lemma 2.6.12 (Flipped Left Cross Contraction)**.**

$$(X \perp\!\!\!\perp Y \mid U, Z) \wedge (Y \perp\!\!\!\perp U \mid Z) \implies Y \perp\!\!\!\perp X, U \mid Z.$$

Proof. By assumption we have the two factorizations:

$$\begin{aligned} K(X, Y, U, Z|T) &= Q(X|U, Z) \otimes K(Y, U, Z|T),\\ K(Y, U, Z|T) &= P(Y|Z) \otimes K(U, Z|T), \end{aligned}$$

with some Markov kernels $Q(X|U, Z)$, $P(Y|Z)$.
Marginalizing out Y in the first equation we get the equality:

$$K(X, U, Z|T) = Q(X|U, Z) \otimes K(U, Z|T).$$

Plugging all three equations into each other we get:

$$\begin{aligned} K(X,Y,U,Z|T) &= Q(X|U,Z) \otimes K(Y,U,Z|T) \\ &= Q(X|U,Z) \otimes P(Y|Z) \otimes K(U,Z|T) \\ &= P(Y|Z) \otimes Q(X|U,Z) \otimes K(U,Z|T) \\ &= P(Y|Z) \otimes K(X,U,Z|T). \end{aligned}$$

In suggestive symbols:

$$K(Y|\cancel{T,X,U},Z) = K(Y|\cancel{T,U},Z).$$

$\square$

Lemma 2.6.13 (T-Restricted Symmetry)**.** *Let $\mathcal{Y}$ and $\mathcal{Z}$ be standard measurable spaces. Then:*

$$X \perp\!\!\!\perp Y \mid Z,T \implies Y \perp\!\!\!\perp X \mid Z,T.$$

Proof. This follows from Flipped Left Cross Contraction with $U = *$ and (Z,T) for Z:

$$(X \perp\!\!\!\perp Y \mid Z,T) \quad \wedge \quad (Y \perp\!\!\!\perp * \mid Z,T) \implies Y \perp\!\!\!\perp X \mid Z,T,$$

together with T-Restricted Right Redundancy:

$$Y \perp\!\!\!\perp * \mid Z,T.$$

In suggestive symbols:

$$K(Y|\cancel{*,T,X},Z) = K(Y|\cancel{T},Z,*).$$

$\square$

2.7. Markov Kernels from Deterministic Mappings

Lemma 2.7.1. *Let $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ be measurable spaces.*

1. *If $f : \mathcal{X} \to \mathcal{Y}$ is measurable then the induced map:*

   $$f_* : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y}), \quad P \mapsto f_*P = (B \mapsto P(f^{-1}(B))),$$

   is measurable as well.

2. *The map:*

   $$\delta : \mathcal{X} \to \mathcal{P}(\mathcal{X}), \quad x \mapsto \delta_x = (A \mapsto \mathbb{1}_A(x)),$$

   is measurable and $\delta^*\mathcal{B}_{\mathcal{P}(\mathcal{X})} = \mathcal{B}_{\mathcal{X}}$. δ is injective iff $\mathcal{B}_{\mathcal{X}}$ separates points.

3. *The map:*

   $$\begin{array}{ccc} \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) & \to & \mathcal{P}(\mathcal{X} \times \mathcal{Y}), \\ (P,Q) & \mapsto & P \otimes Q, \end{array}$$

   is measurable.

4. *If $g : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ is measurable then the map:*

$$\begin{array}{rcl} \mathcal{P}(\mathcal{X}) \times \mathcal{Y} & \to & \mathcal{P}(\mathcal{Z}), \\ (P, y) & \mapsto & g_*(P \otimes \delta_y) \\ & = & (C \mapsto P(\{x \in \mathcal{X} \mid g(x,y) \in C\})) \end{array}$$

is measurable as well.

Remark 2.7.2. *Let $f : \mathcal{Y} \times \mathcal{Z} \to \mathcal{X}$ be measurable and $P(Y) \in \mathcal{P}(\mathcal{Y})$ a fixed probability distribution. Then the map:*

$$K(X|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}, \quad (A, z) \mapsto P(f(Y,z) \in A) =: K(X \in A|Z = z)$$

is a Markov kernel.

Theorem 2.7.3. *Let $\mathcal{Z}$ be any measurable space and $\mathcal{X} = \bar{\mathbb{R}} = [-\infty, \infty]$. Let $K(X|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}$ be a Markov kernel, $P(E)$ be the uniform distribution on $\mathcal{E} := [0,1]$ and:*

$$R(e|z) := \inf \{\tilde{x} \in \mathcal{X} \mid K(X \le \tilde{x}|z) \ge e\},$$

the (conditional) ***quantile function (a.k.a. inverse cumulative distribution function)*** *of $K(X|Z)$. Then we can write $K(X|Z)$ as the push-forward:*

$$K(X|Z) = \delta(R|E, Z) \circ P(E).$$

More explicitly, for $A \in \mathcal{B}_{\mathcal{X}}$ and $z \in \mathcal{Z}$ we have:

$$K(X \in A|Z = z) = P(R(E|z) \in A).$$

Proof. We only need to check the last equation for $A = [-\infty, x]$ and $x \in \bar{\mathbb{R}}$. We then use the following equivalence for $x \in \bar{\mathbb{R}}$, $z \in \mathcal{Z}$ and $e \in [0,1]$, see Lemma 2.7.8:

$$R(e|z) \le x \qquad \Longleftrightarrow \qquad e \le F(x|z),$$

where F is the conditional cumulative distribution function of $K(X|Z)$. So we get:

$$P(R(E|z) \le x) = P(E \le F(x|z)) = F(x|z) := K(X \le x|Z = z).$$

The equality in the middle holds because E is uniformly distributed. This shows the claim. ☐

Remark 2.7.4. *Let $\mathcal{Z}$ be any measurable space and $\mathcal{X}$ be a standard measurable space with a fixed embedding $\iota : \mathcal{X} \hookrightarrow \bar{\mathbb{R}} = [-\infty, \infty]$ onto a Borel subset, which always exists, and $K(X|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}$ a Markov kernel. Then the push-forward Markov kernel:*

$$K(\iota X|Z) : \mathcal{Z} \xrightarrow{K(X|Z)} \mathcal{P}(\mathcal{X}) \xrightarrow{\iota_*} \mathcal{P}(\bar{\mathbb{R}}), \qquad (A, z) \mapsto K(X \in \iota^{-1}(A)|Z = z),$$

satisfies the condition of Theorem 2.7.3. So with those notations we get for all $A \in \mathcal{B}_{\bar{\mathbb{R}}}$ and $z \in \mathcal{Z}$:

$$K(\iota X \in A|Z = z) = P(R(E|z) \in A).$$

Since $\iota(\mathcal{X}) \in \mathcal{B}_{\bar{\mathbb{R}}}$ we get for all $z \in \mathcal{Z}$:

$$0 = K(\iota X \in \iota(\mathcal{X})^\complement | Z = z) = P(R(E|z) \in \iota(\mathcal{X})^\complement).$$

Since R is measurable we get that:

$$D := \{(e,z) \in \mathcal{E} \times \mathcal{Z} \,|\, R(e|z) \in \iota(\mathcal{X})\} \in \mathcal{B}_{\mathcal{E}} \otimes \mathcal{B}_{\mathcal{Z}},$$

with $P(E \in D_z^\complement) = 0$ for all $z \in \mathcal{Z}$. We can then measurably adjust R to get a measurable map:

$$\tilde{R} : \mathcal{E} \times \mathcal{Z} \to \mathcal{X}, \qquad \tilde{R}(e|z) := \iota^{-1}(R(e|z)),$$

for $(e,z) \in D$ and $\tilde{R}(e|z) := \tilde{x}$ for $(e,z) \in D^\complement$ and a fixed point $\tilde{x} \in \mathcal{X}$. With this adjustment we then get for all $A \in \mathcal{B}_{\mathcal{X}}$ and $z \in \mathcal{Z}$:

$$\begin{aligned} K(X \in A|Z = z) &= K(\iota X \in \iota(A)|Z = z) \\ &= P(R(E|z) \in \iota(A)) \\ &= P(\iota(\tilde{R}(E|z)) \in \iota(A)) \\ &= P(\tilde{R}(E|z) \in A), \end{aligned}$$

or in short:

$$K(X|Z) = \delta(\tilde{R}|E, Z) \circ P(E).$$

In other words, Theorem 2.7.3 holds (with those slight adjustments) for all standard measurable spaces $\mathcal{X}$ as well.

In terms of random variables the theorem above states that every distribution Q can be generated by the uniform one $U[0,1]$ and a deterministic map. The theorem below strengthens this claim. It says that every conditional random variable X can be represented in terms of a uniformly distributed random variable E and a measurable map. In short, the above is about 'in distribution' and the one below about 'almost-surely' statements.

Theorem 2.7.5. *Let $\mathcal{X} := \bar{\mathbb{R}}$, $\mathcal{U} := [0,1]$ and $\mathcal{Z}$ any measurable space. Let X, U and Z be conditional random variables taking values in $\mathcal{X}$, $\mathcal{U}$ and $\mathcal{Z}$, resp., such that:*

$$U \underset{K(U,X|Z)}{\perp\!\!\!\perp} X, Z,$$

with $K(U|\cancel{X,Z})$ the uniform distribution on $[0,1]$. Define the interpolated (conditional) cumulative distribution function and its corresponding quantile function via:

$$\begin{aligned} F(x; u|z) &:= K(X < x|Z = z) + u \cdot K(X = x|Z = z), \\ R(e|z) &:= \inf \{\tilde{x} \in \mathcal{X} \,|\, F(\tilde{x}; 1|z) \geq e\}, \end{aligned}$$

and the conditional random variable $E := F(X; U|Z)$. Then we have the (conditional) independence:

$$E \underset{K(U,X|Z)}{\perp\!\!\!\perp} Z,$$

with $K(E|\cancel{Z})$ the uniform distribution on $[0,1]$, and:

$$X = R(E|Z) \quad K(U,X|Z)\text{-a.s.}$$

Proof. After the (joint) measurabilities of F and R are checked the statement directly follows from Lemma 2.7.10 by applying it for every z separately. $\square$

Remark 2.7.6. *With a similar argument as used in Remark 2.7.4 we can in Theorem 2.7.5 replace $\mathcal{X}$ by any standard measurable space. We then use $E := F(\iota X; U|Z)$ to get the conditional independence:*

$$E \underset{K(U,X|Z)}{\perp\!\!\!\perp} Z,$$

with $K(E|Z)$ the uniform distribution on $[0,1]$, and:

$$X = \tilde{R}(E|Z) \quad K(U,X|Z)\text{-a.s.},$$

for some measurable function $\tilde{R}$.

Corollary 2.7.7. *Let X and Z be random variables with values in any standard measurable spaces $\mathcal{X}$ and $\mathcal{Z}$, resp., and with a joint distribution $P(X,Z)$. Then there exists a uniformly distributed random variable E on $[0,1]$ that is P-independent of Z and a measurable function g such that $X = g(E,Z)$ P-almost-surely.*

Proof. The regular conditional probability distribution $P(X|Z)$ exists for standard measurable spaces (and is unique up to a $P(Z)$-zero-set), and is a Markov kernel. Then apply the result from above for $K(X|Z) := P(X|Z)$ to get $g(e,z) := \tilde{R}(e|z)$ and E. $\square$

Proofs - Deterministic Representation of Markov Kernels In this section we generalize a few folklore results via now standard techniques that were introduced in [Dar53, Č82].

Lemma 2.7.8. *Let $\bar{\mathbb{R}} := [-\infty, \infty]$ be endowed with the usual ordering and Borel σ-algebra. Let P be a probability measure on $\bar{\mathbb{R}}$ and $F(x) := P([-\infty, x])$. Then $F : \bar{\mathbb{R}} \to [0,1]$ is non-decreasing, right-continuous with at most countably many discontinuities and $F(\infty) = 1$. So $R(t) := \inf F^{-1}([t,1])$ is a well-defined map $R : [0,1] \to \bar{\mathbb{R}}$, non-decreasing, left-continuous with at most countably many discontinuities and $R(0) = -\infty$. Furthermore, for $x \in \bar{\mathbb{R}}$ and $t \in [0,1]$ we have:*

$$t \le F(x) \iff R(t) \le x.$$

In particular, we have $F(R(t)) \ge t$, thus $R(t) \in F^{-1}([t,1])$ the minimal element. We also have $R(F(x)) \le x$, with equality if and only if $x \in R([0,1])$. Furthermore, F and R are measurable and $R_*\lambda = P$. We also have that R is a reflexive generalized inverse of F, i.e.:

$$F \circ R \circ F = F, \qquad R \circ F \circ R = R.$$

Proof. From the properties of P it is clear that F is non-decreasing, right-continuous and $F(\infty) = 1$.
Let $D_F \subseteq \bar{\mathbb{R}}$ be the set of discontinuities of F and $x \in D_F$. Then there exists a $q(x) \in \mathbb{Q}$ such that $F_-(x) < q(x) < F_+(x)$. If now $x_1 < x_2$ are two such points we get:

$q(x_1) < F_+(x_1) \le F_-(x_2) < q(x_2)$. So the map $q : D_F \to \mathbb{Q}$ is injective. Thus D_F is countable.
Next, we show that $R(t) \in F^{-1}([t,1])$, thus $R(t) = \min F^{-1}([t,1])$. For this let $(x_n)_{n\in\mathbb{N}} \subseteq F^{-1}([t,1])$ be a non-increasing sequence converging to $R(t)$. Then by the right-continuity $F(x_n)$ converges to $F(R(t))$ from above. So we have:

$$F(R(t)) = \inf_{n\in\mathbb{N}} F(x_n) \ge t.$$

It follows that $F(R(t)) \ge t$ and thus $R(t) \in F^{-1}([t,1])$. This shows the claim.
R is clearly non-decreasing, thus has only a countable set of discontinuities $D_R \subseteq [0,1]$ by the same arguments as before, and $R(0) = -\infty$. To see that $R(t)$ is left-continuous let $t \in [0,1]$ and $(t_n)_{n\in\mathbb{N}}$ a non-decreasing sequence converging to t from below. Then by the monotonicity of R we have $\sup_{n\in\mathbb{N}} R(t_n) \le R(t)$. On the other hand we have:

$$t = \sup_{n\in\mathbb{N}} t_n \le \sup_{n\in\mathbb{N}} F(R(t_n)) \le F(\sup_{n\in\mathbb{N}} R(t_n)),$$

implying: $\sup_{n\in\mathbb{N}} R(t_n) \in F^{-1}([t,1])$ and thus $\sup_{n\in\mathbb{N}} R(t_n) \ge R(t)$, leading to equality, which shows the claim.
For any $x \in \bar{\mathbb{R}}$ we have the implication:

$$x \ge R(t) \implies F(x) \ge F(R(t)) \ge t.$$

For any $x \in \bar{\mathbb{R}}$ and any $t \in [0,1]$ we have the implications:

$$\begin{aligned} t \le F(x) &\iff F(x) \in [t,1] \\ &\iff x \in F^{-1}([t,1]) \\ &\implies x \ge \inf F^{-1}([t,1]) = R(t). \end{aligned}$$

Together this shows for any $x \in \bar{\mathbb{R}}$ and $t \in [0,1]$ the equivalence:

$$t \le F(x) \iff R(t) \le x.$$

Since $F(x) \le F(x)$ we get $R(F(x)) \le x$ for all $x \in \bar{\mathbb{R}}$. If equality holds then $x \in R([0,1])$. And, if $x = R(t)$ for some $t \in [0,1]$ then we use the inequalities $x \ge R(F(x))$ and $F(R(t)) \ge t$ to conclude:

$$x \ge R(F(x)) = R(F(R(t))) \ge R(t) = x,$$

showing equality, and that:

$$R \circ F \circ R = R.$$

Similarly for $t = F(x)$ we get:

$$t \le F(R(t)) = F(R(F(x))) \le F(x) = t,$$

showing

$$F \circ R \circ F = F.$$

Now consider the uniform distribution λ on $[0,1]$ and any $x \in \bar{\mathbb{R}}$. Then we have:

$$\begin{array}{rcl} (R_*\lambda)([-\infty,x]) & = & \lambda(R^{-1}([-\infty,x])) \\ & = & \lambda(t \in [0,1] \,|\, R(t) \leq x) \\ & = & \lambda(t \in [0,1] \,|\, t \leq F(x)) \\ & = & \lambda([0,F(x)]) \\ & = & F(x) \\ & = & P([-\infty,x]). \end{array}$$

It follows that: $R_*\lambda = P$. □

Lemma 2.7.9. *Let the notation be like in lemma 2.7.8. For $u \in [0,1]$ and $x \in \bar{\mathbb{R}}$ define:*

$$F_u(x) := E(x;u) := P([-\infty,x)) + u \cdot P(\{x\}).$$

Then $E : \bar{\mathbb{R}} \times [0,1] \to [0,1]$ is measurable, non-decreasing in both arguments with $F_0(-\infty) = 0$, $F_1(\infty) = 1$, F_0 is left-continuous and

$$F_u(\tilde{x}) \leq F_1(\tilde{x}) \leq F_0(x) \leq F_u(x)$$

for any $\tilde{x} < x$, $u \in [0,1]$. We further have for every $u \in (0,1]$:

$$R \circ F_u \circ R = R,$$

and $R \circ F_u = \mathrm{id}_{\bar{\mathbb{R}}}$ P-almost-surely for any $u \in (0,1]$.

Proof. Most of the properties are clear from its definition. Let $\tilde{x} < x$ then $[-\infty,\tilde{x}] \subseteq [-\infty,x)$ and thus $F_1(\tilde{x}) \leq F_0(x)$.
To show $R \circ F_u \circ R = R$ fix a $t \in [0,1]$, $u \in (0,1]$ and let $x := R(t)$. If F_1 is continuous in x then $F_u = F_1$ and the claim $R \circ F_1 \circ R = R$ was already shown using the inequalities:

$$x \geq R(F_1(x)) = R(F_1(R(t))) \geq R(t) = x.$$

So let us assume that F_1 is discontinuous in $x = R(t)$. Then $F_u(x) \in (F_0(x), F_1(x)]$. We have:

$$R(F_u(x)) = \min\{\tilde{x} \in \bar{\mathbb{R}} \,|\, F_1(\tilde{x}) \geq F_u(x)\}.$$

If $F_1(\tilde{x}) \geq F_u(x) > F_0(x)$ then $\tilde{x} \geq x$, otherwise $\tilde{x} < x$ leads to the contradiction $F_1(\tilde{x}) \leq F_0(x)$. Since clearly $F_1(x) \geq F_u(x)$ we must have:

$$R(F_u(x)) = x,$$

with $x = R(t)$, which proves the claim: $R \circ F_u \circ R = R$ for $u \in (0,1]$.
We now want to show that $R \circ F_u = \mathrm{id}_{\bar{\mathbb{R}}}$ P-a.s. for $u \in (0,1]$. From $R \circ F_u \circ R = R$ we already see that $R \circ F_u|_{R([0,1])} = \mathrm{id}_{R([0,1])}$. We will see below that $C := \bar{\mathbb{R}} \setminus R([0,1])$ is measurable and $P(C) = 0$, which will prove the claim.
In the following we will only need $F = F_1$. First, by lemma 2.7.8 we know that for any

$x \in \bar{\mathbb{R}}$ we have $R(F(x)) \leq x$ with equality if and only if $x \in R([0,1])$. So this gives us the equivalence:

$$x \in C \iff x > R(F(x)).$$

We now claim that $(R(F(x)), x] \subseteq C$ for every $x \in C$: Indeed, If $\tilde{x} \in (R(F(x)), x]$ then:

$$F(x) = F(R(F(x))) \leq F(\tilde{x}) \leq F(x)$$

and thus $F(\tilde{x}) = F(x)$, from which follows that $R(F(\tilde{x})) = R(F(x)) < \tilde{x}$ and ergo $\tilde{x} \in C$.

It follows that C is the union of such intervals $(R(F(x)), x]$ with $x \in C$. Furthermore, $F(C)$ is contained in the set of discontinuities D_R of R: otherwise there would be an $x \in C$ and a $t \geq F(x)$ such that $R(t) \in (R(F(x)), x] \subseteq C$, which is a contradiction. Since D_R is countable it must follow that $F(C)$ and thus also $R(F(C))$ is at most countable. Write $R(F(C)) = \{x_n \,|\, n \in \mathbb{N}\}$, which is the set of the possible left end-points of the above intervals. For each fixed $n \in \mathbb{N}$ let

$$C_n := \{x \in C \,|\, R(F(x)) = x_n\},$$

which is, as a union of intervals $(x_n, x]$, $x \in C_n$, either of the form $(x_n, \bar{x}_n]$ or $(x_n, \bar{x}_n)$ with $\bar{x}_n := \sup C_n$. In both cases we can cover C_n by $C_{n,m} := (x_n, x_{n,m}]$ with $x_{n,m} \in C_n$ either equal to $\bar{x}_n$ or converging to it from below for running m. So we can write C as the countable union:

$$C = \bigcup_{n,m \in \mathbb{N}} C_{n,m}.$$

We now have for each $x = x_{n,m}$:

$$P(C_{n,m}) = P((x_n, x]) = P((R(F(x)), x]) = F(x) - F(R(F(x))) = F(x) - F(x) = 0.$$

This implies:

$$P(C) = P\left(\bigcup_{n,m \in \mathbb{N}} C_{n,m}\right) \leq \sum_{n,m \in \mathbb{N}} P(C_{n,m}) = 0,$$

showing that $P(C) = 0$ and thus:

$$R \circ F_u = \mathrm{id}_{\bar{\mathbb{R}}} \qquad P\text{-a.s.}$$

for $u \in (0,1]$. □

Lemma 2.7.10. *Let the notations be like in lemma 2.7.8 and lemma 2.7.9. Let λ be the uniform distribution on $[0,1]$ and $\bar{P} := P \otimes \lambda$ the product distribution on $\bar{\mathbb{R}} \times [0,1]$. For every $e \in [0,1]$ define the event:*

$$\{E \leq e\} := \{(x,u) \in \bar{\mathbb{R}} \times [0,1] \mid E(x;u) \leq e\}.$$

Then $\bar{P}(E \leq e) = e$. In other words, the random variable:

$$\begin{array}{rcl} E \;:\; \bar{\mathbb{R}} \times [0,1] & \to & [0,1], \\ (x,u) & \mapsto & P([-\infty, x)) + u \cdot P(\{x\}), \end{array}$$

is uniformly distributed under $\bar{P} = P \otimes \lambda$.
Furthermore, $R(E) = X$ $\bar{P}$-a.s., where $X : \bar{\mathbb{R}} \times [0,1] \to \bar{\mathbb{R}}$ is the canonical projection onto the first factor: $X(x,u) := x$, which has distribution P.

Proof. First, since $\lambda(\{0\}) = 0$ we can w.l.o.g. exclude $u = 0$ and restrict $\bar{P}$ to $\bar{\mathbb{R}} \times (0,1]$. We have seen in lemma 2.7.9 that $R \circ F_u \circ R = R$ for $u \in (0,1]$, which translates to:

$$R \circ E|_{R([0,1]) \times (0,1]} = X|_{R([0,1]) \times (0,1]}.$$

Also with $C := \bar{\mathbb{R}} \setminus R([0,1])$ we get:

$$\bar{P}(C \times (0,1]) = P(C) \cdot \lambda((0,1]) = 0 \cdot 1 = 0.$$

So we get the second claim that:

$$R \circ E = X \qquad \bar{P}\text{-a.s.}$$

Now we turn to $\{E \leq e\}$ for $e \in [0,1]$. We abbreviate $U : \bar{\mathbb{R}} \times [0,1] \to [0,1]$ to be the projection onto the second factor: $U(x,u) := u$, which is uniformly distributed under $\bar{P}$, and also $p(x) := P(\{x\}) = F_1(x) - F_0(x)$. With these notations: $E = F_0(X) + U \cdot p(X)$. First, we show that $\bar{P}(E = e) = 0$ for all $e \in [0,1]$. For this let $x := R(e)$. Then by the above ($R(E) = X$ $\bar{P}$-a.s.) we have:

$$\bar{P}(E = e) = \bar{P}(E = e,\, X = x).$$

We have to distinguish between two cases: $p(x) = 0$ and $p(x) > 0$.
Case $p(x) = 0$: We have:

$$\begin{aligned} \bar{P}(E = e) &= \bar{P}(E = e, X = x) \\ &\leq \bar{P}(X = x) \\ &= p(x) \\ &= 0. \end{aligned}$$

Case $p(x) > 0$: We get:

$$\begin{aligned} \bar{P}(E = e) &= \bar{P}(E = e, X = x) \\ &= \bar{P}(F_0(X) + U \cdot p(X) = e,\, X = x) \\ &= \bar{P}\left(U = \frac{e - F_0(x)}{p(x)},\, X = x\right) \\ &= \lambda\left(\left\{\frac{e - F_0(x)}{p(x)}\right\}\right) \cdot p(x) \\ &= 0. \end{aligned}$$

To prove $\bar{P}(E \leq e) = e$ for $e \in [0,1]$ we have several cases:
Case $e \in F_1(\bar{\mathbb{R}})$: Let $\tilde{x}$ be any element in $\bar{\mathbb{R}}$ with $e = F_1(\tilde{x})$ (e.g. $\tilde{x} = R(e)$). Then we

get:

$$\begin{aligned}
\bar{P}(E \leq e) &= \bar{P}(E \leq F_1(\tilde{x})) \\
&= \bar{P}(R(E) \leq \tilde{x}) \\
&\overset{R \circ E = X}{=} \bar{P}(X \leq \tilde{x}) \\
&= P([-\infty, \tilde{x}]) \cdot \lambda((0,1]) \\
&= F_1(\tilde{x}) \cdot 1 \\
&= e.
\end{aligned}$$

For the cases $e \notin F_1(\bar{\mathbb{R}})$ we put $x := R(e)$ and $\tilde{e} := F_0(x)$.
Then by definition, x is minimal with $F_1(x) \geq e$. We also have $\tilde{e} = F_0(x) \leq e$. Otherwise: $e < F_0(x) = \sup_{\tilde{x}<x} F_1(\tilde{x})$ implied that there existed $\tilde{x} < x$ with $e < F_1(\tilde{x}) \leq F_0(x)$, which is a contradiction to the minimality of $x = R(e)$. Since $\tilde{e} \leq e$ we can decompose:

$$\bar{P}(E \leq e) = \bar{P}(E < \tilde{e}) + \bar{P}(E = \tilde{e}) + \bar{P}(\tilde{e} < E \leq e).$$

We have already seen that the second term $\bar{P}(E = \tilde{e}) = 0$ vanishes.
For the first term we have:

$$\begin{aligned}
\bar{P}(E < \tilde{e}) &= \bar{P}(E < F_0(x)) \\
&= \bar{P}(E < \sup_{\tilde{x}<x} F_1(\tilde{x})) \\
&= \sup_{\tilde{x}<x} \bar{P}(E \leq F_1(\tilde{x})) \\
&\overset{(*)}{=} \sup_{\tilde{x}<x} F_1(\tilde{x}) \\
&= F_0(x) \\
&= \tilde{e}.
\end{aligned}$$

Equation (*) comes from the previous case for $F_1(\tilde{x}) \in F_1(\bar{\mathbb{R}})$.
For the third term $\bar{P}(\tilde{e} < E \leq e)$ first note that $E \in (\tilde{e}, e]$ implies that $X = x$ $\bar{P}$-a.s. by applying R: Indeed, every element $t \in (\tilde{e}, e] \subseteq (F_0(x), F_1(x)]$ can be written as $t = F_{\tilde{u}}(x)$ for an $\tilde{u} \in (0,1]$ and we can use:

$$R(t) = R(F_{\tilde{u}}(R(e)) = R(e) = x.$$

For $p(x) > 0$ and the above we get:

$$\begin{aligned}
\bar{P}(\tilde{e} < E \leq e) &= \bar{P}(\tilde{e} < E \leq e,\, X = x) \\
&= \bar{P}(0 < F_0(X) + U \cdot p(X) - F_0(x) \leq e - \tilde{e}, X = x) \\
&= \bar{P}(0 < U \leq \frac{e - \tilde{e}}{p(x)}, X = x) \\
&= \lambda\left(\left(0, \frac{e - \tilde{e}}{p(x)}\right]\right) \cdot P(\{x\}) \\
&= \frac{e - \tilde{e}}{p(x)} \cdot p(x) \\
&= e - \tilde{e}.
\end{aligned}$$

For the case $p(x) = 0$, $\bar{P}(\tilde{e} < E \leq e, X = x)$ can be upper bounded by $\bar{P}(X = x) = p(x) = 0$ as before, but we also have $\tilde{e} - e = 0$ in this case, and the equality stays trivially true as well.
Putting all together we get:

$$\begin{aligned}\bar{P}(E \leq e) &= \bar{P}(E < \tilde{e}) + \bar{P}(E = \tilde{e}) + \bar{P}(\tilde{e} < E \leq e) \\ &= \tilde{e} + 0 + e - \tilde{e} \\ &= e.\end{aligned}$$

This shows the claim. $\square$

3. Graph Theory

3.1. Core Concepts

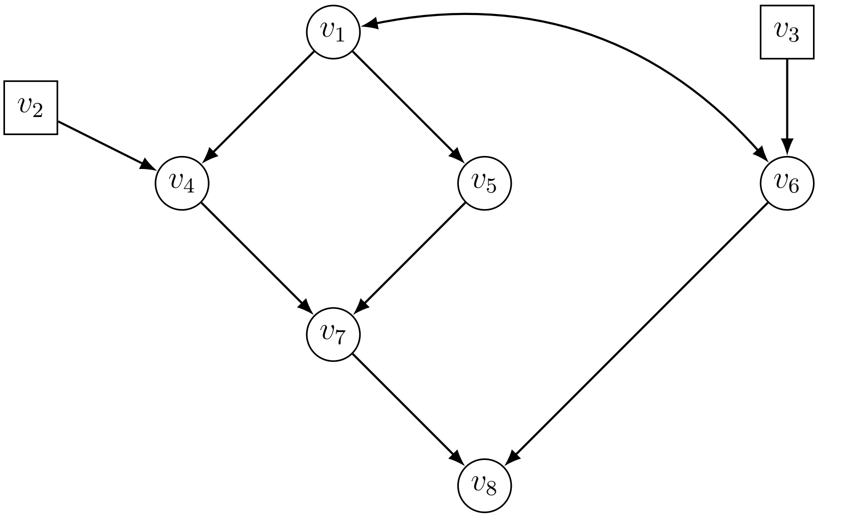


Figure 3: Conditional Acyclic Directed Mixed Graph (CADMG).

Definition 3.1.1 (Conditional directed mixed graphs (CDMG))**.** *A **conditional directed mixed graph (CDMG)** G—per definition—consists of two (disjoint) sets of vertices (also called nodes):*

i.) J, whose elements are called input nodes,

ii.) V, whose elements are called output nodes,

and two (disjoint) sets of edges:

iii.) $E \subseteq (J \cup V) \times V$ the set of directed edges,

iv.) $L \subseteq V \times V/((v_1, v_2) \sim (v_2, v_1))$, the set of bi-directed edges,

with: $(v_1, v_2) \in L \implies v_1 \neq v_2 \wedge (v_2, v_1) \in L$.

Notation 3.1.2. *Let $G = (J, V, E, L)$ be a CDMG. We will write:*

1. $v \in G$ to mean $v \in J \cup V$,

2. $v_1 \longrightarrow v_2 \in G$ to mean $(v_1, v_2) \in E$,

3. $v_1 \longleftarrow v_2 \in G$ to mean $(v_2, v_1) \in E$,

4. $v_1 \longleftrightarrow v_2 \in G$ to mean $(v_1, v_2) \in L$,

5. $v_1 \ast\!\!\longrightarrow v_2 \in G$ to mean that either $v_1 \longrightarrow v_2 \in G$ or $v_1 \longleftrightarrow v_2 \in G$,

6. $v_1 \mathrel{\leftarrow\!\!\ast} v_2 \in G$ to mean that either $v_1 \leftarrow v_2 \in G$ or $v_1 \leftrightarrow v_2 \in G$,

7. $v_1 \mathrel{\ast\!\!-\!\!\ast} v_2 \in G$ to mean that either $v_1 \rightarrow v_2 \in G$ or $v_1 \leftarrow v_2 \in G$ or $v_1 \leftrightarrow v_2 \in G$.

The star stands for a placeholder to mean: "arrowhead or tail".

Definition 3.1.3. *Let $G = (J, V, E, L)$ be a CDMG.*

1. *If $v_1 \mathrel{\ast\!\!-\!\!\ast} v_2 \in G$ then we call v_1 and v_2* ***adjacent in*** *G.*
2. *Edges of the form $v_1 \leftarrow v_2$ or $v_1 \leftrightarrow v_2$ are called* ***into*** *v_1.*
   Edges of the form $v_1 \rightarrow v_2$ or $v_1 \leftrightarrow v_2$ are called ***into*** *v_2.*
3. *Edges of the form $v_1 \rightarrow v_2$ or $v_2 \leftarrow v_1$ are called* ***out of*** *v_1.*

Remark 3.1.4. *With the notations 3.1.2 the restrictions in definition 3.1.1 mean that the nodes $j \in J$ will not have any arrowheads pointing towards them: $j \mathrel{\leftarrow\!\!\ast} v \notin G$. Nodes $j \in J$ can only point towards nodes $v \in V$: edges $j \rightarrow v$ are allowed. Furthermore, no two nodes in J are adjacent.*

Definition 3.1.5 (Walks)**.** *Let $G = (J, V, E, L)$ be a CDMG and $v, w \in G$.*

1. *A* ***walk*** *from v to w in G is a finite alternating sequence of adjacent nodes and edges*

   $$v = v_0, a_0, v_1, \dots v_{n-1}, a_{n-1}, v_n = w$$

   in G for some $n \geq 0$, i.e. such that for every $k = 0, \dots, n-1$ we have that $a_k = (v_k, v_{k+1}) \in E \cup L$ or $a_k = (v_{k+1}, v_k) \in E$, and with end nodes $v_0 = v$ and $v_n = w$. An example walk from v_0 to v_3 could look like:

   $$v_0 \rightarrow v_1 \leftarrow v_2 \leftrightarrow v_3, \qquad \text{with} \qquad v_0 \rightarrow v_1, v_2 \rightarrow v_1 \in E,\ v_2 \leftrightarrow v_3 \in L.$$

   The same node may appear multiple times in a walk. Also the ***trivial walk*** *consisting of a single node $v_0 \in G$ is allowed (if $v = w$). The walk is called* ***into*** *v_0 if $a_0 = v_0 \mathrel{\leftarrow\!\!\ast} v_1$, and* ***out of*** *$v_0$ if $a_0 = v_0 \rightarrow v_1$. Similarly, it is called* ***into*** *v_n if $a_{n-1} = v_{n-1} \mathrel{\ast\!\!\rightarrow} v_n$ and* ***out of*** *v_n if $a_{n-1} = v_{n-1} \leftarrow v_n$.*

2. *A* ***directed walk*** *from v to w in G is of the form:*

   $$v = v_0 \rightarrow v_1 \rightarrow \cdots \rightarrow v_{n-1} \rightarrow v_n = w,$$

   for some $n \geq 0$, where all arrowheads point in the direction of w and there are no arrowheads pointing back.

3. *A* ***bi-directed walk*** *from v to w in G is of the form:*

   $$v = v_0 \leftrightarrow v_1 \leftrightarrow \cdots \leftrightarrow v_{n-1} \leftrightarrow v_n = w,$$

   for some $n \geq 0$, where all edges are bi-directed.

4. *A **collider walk** from v to w in G is of the form:*

$$v = v_0 \ast\!\!\rightarrow v_1 \leftrightarrow \cdots \leftrightarrow v_{n-1} \leftarrow\!\!\ast\, v_n = w,$$

for some $n \geq 0$, where all nodes in between v and w have two arrowheads pointing towards them (a.k.a. collider). Note that for $n = 1$ this reads: $v \ast\!\!-\!\!\ast\, w \in G$.

5. *A walk is called **path** if no node occurs more than once.*

6. *A **bifurcation** between v and w in G is a walk of the form:*

$$v = v_0 \leftarrow v_1 \leftarrow \cdots \leftarrow v_{k-1} \leftarrow\!\!\ast\, v_k \rightarrow \cdots \rightarrow v_{n-1} \rightarrow v_n = w,$$

*such that $v \neq w$, the walk contains both endnodes exactly once, every node has at most one arrowhead pointing towards it, and both endnodes have exactly one arrowhead pointing towards them. If the edge $v_{k-1} \leftarrow\!\!\ast\, v_k$ is directed ($v_{k-1} \leftarrow v_k$) then we say that the bifurcation has **source** v_k.*

Definition 3.1.6 (Family relationships)**.** *Let $G = (J, V, E, L)$ be a CDMG, $v, w \in V$ and $A \subseteq J \cup V$ a subset of nodes. We then define:*

1. *The set of **parents** of v in G:*

$$\mathrm{Pa}^G(v) := \{w \in G \,|\, w \rightarrow v \in G\}.$$

*The set of **parents** of A in G:*

$$\mathrm{Pa}^G(A) := \bigcup_{v \in A} \mathrm{Pa}^G(v).$$

2. *The set of **children** of v in G:*

$$\mathrm{Ch}^G(v) := \{w \in G \,|\, v \rightarrow w \in G\}.$$

*The set of **children** of A in G:*

$$\mathrm{Ch}^G(A) := \bigcup_{v \in A} \mathrm{Ch}^G(v).$$

3. *The set of **siblings** of v in G:*

$$\mathrm{Sib}^G(v) := \{w \in G \,|\, v \leftrightarrow w \in G\}.$$

4. *The set of **ancestors** of v in G:*

$$\mathrm{Anc}^G(v) := \{w \in G \,|\, \exists \text{ directed walk: } w \rightarrow \cdots \rightarrow v \in G\}.$$

Note: $v \in \mathrm{Anc}^G(v)$.

*The set of **ancestors** of A in G:*

$$\mathrm{Anc}^G(A) := \bigcup_{v \in A} \mathrm{Anc}^G(v).$$

Note: $A \subseteq \mathrm{Anc}^G(A)$.

5. *The set of **descendants** of v in G:*

$$\mathrm{Desc}^G(v) := \{w \in G \,|\, \exists \textit{ directed walk: } v \longrightarrow \cdots \longrightarrow w \in G\}.$$

Note: $v \in \mathrm{Desc}^G(v)$.

*The set of **descendants** of A in G:*

$$\mathrm{Desc}^G(A) := \bigcup_{v \in A} \mathrm{Desc}^G(v).$$

Note: $A \subseteq \mathrm{Desc}^G(A)$.

6. *The set of **non-descendants** of A in G:*

$$\mathrm{NonDesc}^G(A) := (J \cup V) \setminus \mathrm{Desc}^G(A).$$

7. *The **strongly connected component** of v in G:*

$$\mathrm{Sc}^G(v) := \mathrm{Anc}^G(v) \cap \mathrm{Desc}^G(v).$$

Note: $v \in \mathrm{Sc}^G(v)$.

*The (union of) **strongly connected components** of A in G:*

$$\mathrm{Sc}^G(A) := \bigcup_{v \in A} \mathrm{Sc}^G(v).$$

Note: $A \subseteq \mathrm{Sc}^G(A)$.

8. *The **district** of v in G:*

$$\mathrm{Dist}^G(v) := \{w \in G \,|\, \exists \textit{ bi-directed walk: } v \leftrightarrow v_1 \leftrightarrow \cdots \leftrightarrow v_{n-1} \leftrightarrow w \in G\}.$$

Note: $v \in \mathrm{Dist}^G(v)$.

*The **district** of A in G:*

$$\mathrm{Dist}^G(A) := \bigcup_{v \in A} \mathrm{Dist}^G(v).$$

Note: $A \subseteq \mathrm{Dist}^G(A)$.

Definition 3.1.7 (Acyclicity)**.** *A CDMG $G = (J, V, E, L)$ is called **acyclic** if there does not exist any non-trivial directed walk from v to itself in G for any node $v \in G$.*

Definition 3.1.8. *A Conditional Directed Mixed Graph (CDMG) $G = (J, V, E, L)$ is called:*

1. *Conditional Acyclic Directed Mixed Graph (CADMG) if G is acyclic.*
2. *Directed Mixed Graph (DMG) if $J = \emptyset$.*
3. *Acyclic Directed Mixed Graph (ADMG) if G is acyclic and $J = \emptyset$.*
4. *Conditional Directed Graph (CDG) if $L = \emptyset$.*
5. *Directed Graph (DG) if $J = \emptyset$ and $L = \emptyset$.*
6. *Conditional Directed Acyclic Graph (CDAG) if G is acyclic and $L = \emptyset$.*
7. *Directed Acyclic Graph (DAG) if G is acyclic, $J = \emptyset$ and $L = \emptyset$.*

Definition 3.1.9 (Topological order)**.** *Let $G = (J, V, E, L)$ be a CDMG. A **topological order** of G is a total order $<$ of $J \cup V$ such that for all $v, w \in G$:*

$$v \in \mathrm{Pa}^G(w) \implies v < w.$$

Equivalently, it can be described as an indexing of the nodes $J \cup V = \{v_1, \dots, v_K\}$ where parents always precede their children.

Lemma 3.1.10. *A CDMG $G = (J, V, E, L)$ is acyclic if and only if it has a topological order.*

Definition 3.1.11 (Predecessors)**.** *Let $G = (J, V, E, L)$ be a CDMG and $<$ a total order of $J \cup V$. The set of **predecessors** of v in G are:*

$$\mathrm{Pred}^G_<(v) := \{w \in G \,|\, w < v\}.$$

We also put:

$$\mathrm{Pred}^G_\leq(v) := \{w \in G \,|\, w < v\} \cup \{v\}.$$

3.2. Operations on Graphs

3.2.1. Hard Interventions on Graphs

Definition 3.2.1 (Hard intervention on CDMGs)**.** *Let $G = (J, V, E, L)$ be a CDMG and $W \subseteq J \cup V$ a subset of nodes. The **intervened CDMG** w.r.t. W of G is the CDMG:*

$$G_{\mathrm{do}(W)} := (J_{\mathrm{do}(W)}, V_{\mathrm{do}(W)}, E_{\mathrm{do}(W)}, L_{\mathrm{do}(W)}),$$

where:

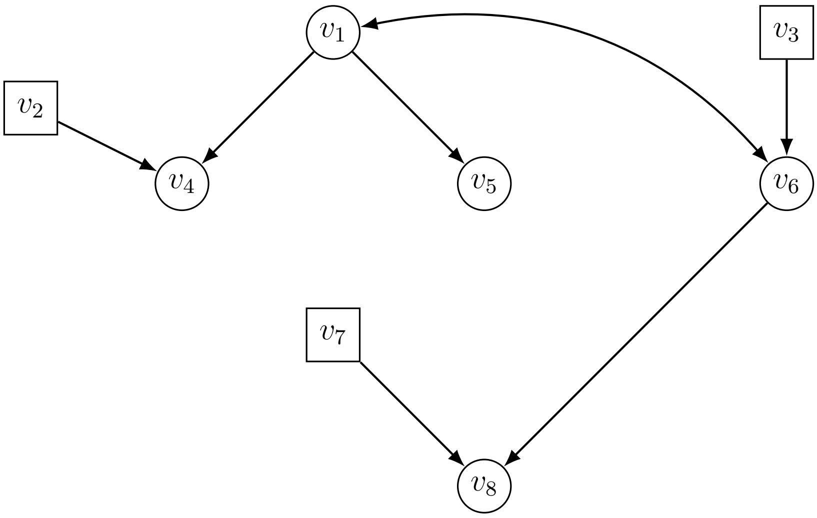


Figure 4: The CADMG from Figure 3 after hard intervention on node v_7.

i.) $J_{\mathrm{do}(W)} := J \cup W$,

ii.) $V_{\mathrm{do}(W)} := V \setminus W$,

iii.) $E_{\mathrm{do}(W)} := E \setminus \{v \longrightarrow w \,|\, v \in G, w \in W\}$,

iv.) $L_{\mathrm{do}(W)} := L \setminus \{v \longleftrightarrow w \,|\, v \in G, w \in W\}$,

where we turn all nodes from W into input nodes and remove all edges into nodes from W.

Remark 3.2.2. *If G is acyclic then also $G_{\mathrm{do}(W)}$ is acyclic and a topological order for G is also one for $G_{\mathrm{do}(W)}$.*

Lemma 3.2.3 (Hard interventions commute)**.** *Let $G := (J, V, E, L)$ be a CDMG and $W_1, W_2 \subseteq J \cup V$ two subsets of nodes from G. Then we have:*

$$\left(G_{\mathrm{do}(W_1)}\right)_{\mathrm{do}(W_2)} = \left(G_{\mathrm{do}(W_2)}\right)_{\mathrm{do}(W_1)} = G_{\mathrm{do}(W_1 \cup W_2)}.$$

The following proposition expresses the existence of a bifurcation with a source in terms of ancestral relations in intervened graphs.

Proposition 3.2.4. *Let $G = (J, V, E, L)$ be a CDMG. For $v, w, c \in V \cup J$: there exists a bifurcation between v and w in G with source c if and only if $v \neq w$ and $c \in \mathrm{Anc}^{G_{\mathrm{do}(w)}}(v) \setminus \{v\}$ and $c \in \mathrm{Anc}^{G_{\mathrm{do}(v)}}(w) \setminus \{w\}$.*

Proof. A bifurcation between v and w with source c is a walk in G of the form $v \longleftarrow \cdots \longleftarrow c \longrightarrow \cdots \longrightarrow w$, where both v and w appear exactly once on the walk. This shows "$\implies$". For the other implication, note that $c \in \mathrm{Anc}^{G_{\mathrm{do}(w)}}(v) \setminus \{v\}$ implies that there is a non-trivial directed path from c to v that does not pass through w. Similarly, $c \in \mathrm{Anc}^{G_{\mathrm{do}(v)}}(w) \setminus \{w\}$ implies that there is a non-trivial directed path from c to w that does not pass through v. The concatenation of the two paths $v \longleftarrow \cdots \longleftarrow c \longrightarrow \cdots \longrightarrow w$ is then a bifurcation between v and w with source c. $\square$

3.2.2. Node Splitting on Graphs

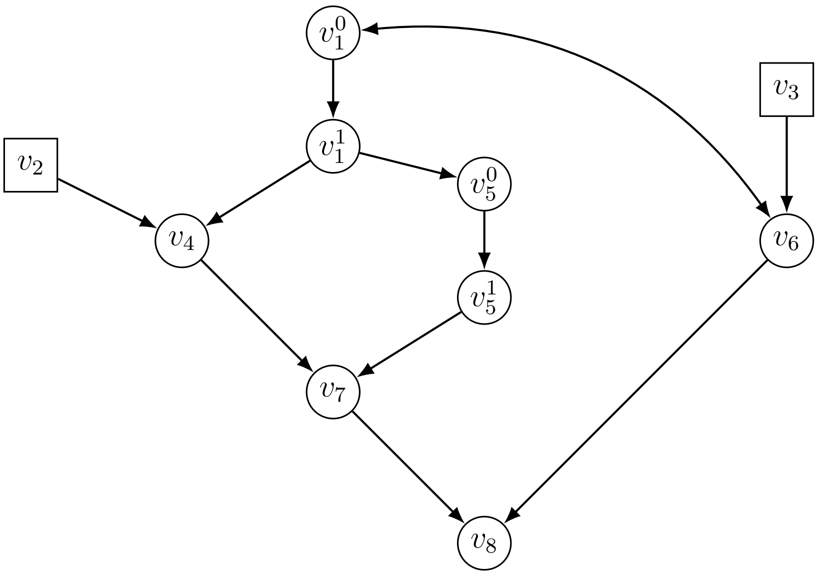


Figure 5: The CADMG from Figure 3 after node-splitting v_1 and v_5.

In this subsection we introduce the operation of *node-splitting*. This is helpful whenever we want to distinguish between two versions of the same variable that in the absence of an intervention share the same value. This is useful to model a property that is constant over some time interval (in the absence of an intervention targeting that property). Conceptually, this comes in handy when we measure the value of a variable right before we intervene on it. The node-splitting operation can be used to represent "single-world" counterfactuals.

Definition 3.2.5 (Node-splitting on CDMGs)**.** *Let $G = (J, V, E, L)$ be a CDMG and $W \subseteq V$ a subset of the output nodes. The* ***node-split graph*** *w.r.t. W of G is the CDMG:*

$$G_{\mathrm{split}(W)} := \left(J_{\mathrm{split}(W)}, V_{\mathrm{split}(W)}, E_{\mathrm{split}(W)}, L_{\mathrm{split}(W)}\right),$$

constructed as follows. We first make two disjont copies of the nodes in W:

$$W^0 := \left\{w^0 \,\middle|\, w \in W\right\}, \qquad W^1 := \left\{w^1 \,\middle|\, w \in W\right\}.$$

Note that we consider $w^0 \neq w^1$ for $w \in W$. Additionally (for convenience), for $v \in J \cup V \setminus W$ we put:

$$v^0 := v^1 := v.$$

We then define:

i.) $J_{\mathrm{split}(W)} := J$,

ii.) $V_{\mathrm{split}(W)} := (V \setminus W) \,\dot\cup\, W^0 \,\dot\cup\, W^1$, *iii.)* $E_{\mathrm{split}(W)} := \{v_1^1 \to v_2^0 \,|\, v_1 \to v_2 \in E\} \cup \{w^0 \to w^1 \,|\, w \in W\}$,

iv.) $L_{\mathrm{split}(W)} := \{v_1^0 \leftrightarrow v_2^0 \,|\, v_1 \leftrightarrow v_2 \in L\}$.

So all incoming edges onto nodes in W become incoming edges into the corresponding nodes in W^0, all outgoing edges out of nodes in W become outgoing edges out of the corresponding nodes in W^1, and edges $w^0 \to w^1$ are added for all nodes in W.

Remark 3.2.6. *For a CADMG $G = (J, V, E, L)$, also $G_{\mathrm{split}(W)}$ is acyclic. If $<$ is any topological order of G given by enumerating all nodes $v \in J \cup V$ via:*

$$v_1 < v_2 < \cdots < v_n,$$

then, for instance, a topological order for $G_{\mathrm{split}(W)}$ can be achieved by assigning for a node $v_j \in W$ with index j the node v_j^0 the index $j - \frac{1}{3}$ and v_j^1 the index $j + \frac{1}{3}$, and then ordering all nodes according to their index value.

Lemma 3.2.7 (Two disjoint node-splittings commute). *Let $G = (J, V, E, L)$ be a CDMG and $W_1, W_2 \subseteq V$ two disjoint subsets of the output nodes of G. Then the CDMG obtained from first node-splitting W_1 and then node-splitting W_2 is the same CADMG that arises from first node-splitting W_2 and then node-splitting W_1:*

$$\left(G_{\mathrm{split}(W_1)}\right)_{\mathrm{split}(W_2)} = \left(G_{\mathrm{split}(W_2)}\right)_{\mathrm{split}(W_1)} = G_{\mathrm{split}(W_1 \dot\cup W_2)}.$$

Lemma 3.2.8 (Disjoint hard interventions and node-splittings commute). *Let $G = (J, V, E, L)$ be a CDMG and $W_1 \subseteq J \cup V$ and $W_2 \subseteq V$ two disjoint subsets of nodes of G. Then the CDMG obtained from first hard intervening on W_1 and then node-splitting W_2 is the same CDMG that arises from first node-splitting W_2 and then hard intervening on W_1.*

$$\left(G_{\mathrm{do}(W_1)}\right)_{\mathrm{split}(W_2)} = \left(G_{\mathrm{split}(W_2)}\right)_{\mathrm{do}(W_1)}.$$

3.2.3. Node Splitting Hard Interventions on Graphs

In this subsection we introduce *node-splitting hard interventions*. They are simply a node-splitting operation (splitting each $w \in W$ into $w^0 \to w^1$) followed by a hard interventions on the "later" nodes W^1. They can represent *single-world intervention graphs (SWIGs)*, which model the same output variable both before and after a hard intervention, see [RR13a, RR13b].

Definition 3.2.9 (Node-splitting hard intervention on CADMGs). *Let $G = (J, V, E, L)$ be a CADMG and $W \subseteq V$ a subset of the output nodes. The* ***single-world intervention graph (SWIG)*** *w.r.t. W of G is the CADMG:*

$$G_{\mathrm{swig}(W)} := \left(J_{\mathrm{swig}(W)}, V_{\mathrm{swig}(W)}, E_{\mathrm{swig}(W)}, L_{\mathrm{swig}(W)}\right),$$

constructed as follows. We first make two disjont copies of the nodes in W:

$$W^o := \{w^o \,|\, w \in W\}, \qquad W^i := \left\{w^i \,\middle|\, w \in W\right\}.$$

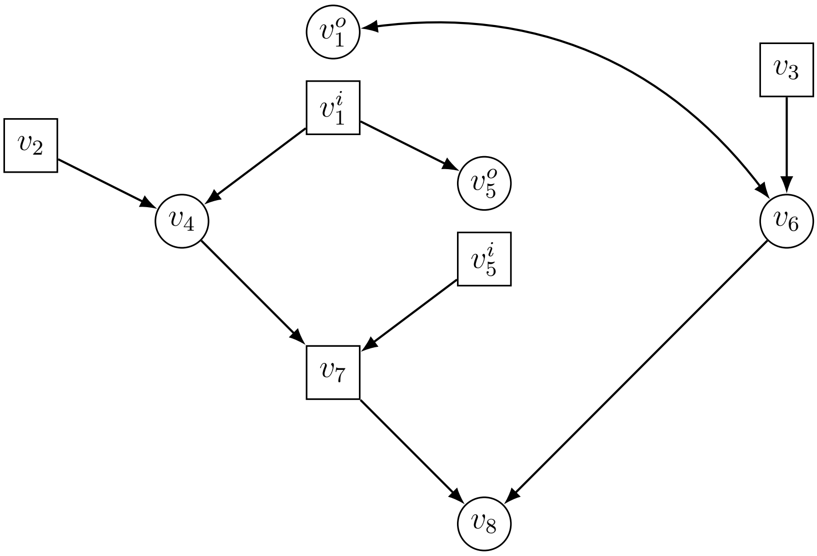


Figure 6: The CADMG from Figure 3 after a node-splitting hard intervention on v_1 and v_5.

Note that we consider $w^o \neq w^i$ for $w \in W$. However, for brevity, for $v \in J \cup V \setminus W$ we put:

$$v^o := v^i := v.$$

We then define:

i.) $J_{\mathrm{swig}(W)} := J \,\dot{\cup}\, W^i$,

ii.) $V_{\mathrm{swig}(W)} := (V \setminus W) \,\dot{\cup}\, W^o$,

iii.) $E_{\mathrm{swig}(W)} := \{v_1^i \longrightarrow v_2^o \,|\, v_1 \longrightarrow v_2 \in E\}$,

iv.) $L_{\mathrm{swig}(W)} := \{v_1^o \longleftrightarrow v_2^o \,|\, v_1 \longleftrightarrow v_2 \in L\}$.

where we turn all nodes of W^i into input nodes, removing all edges into W^i, and we turn all nodes of W^o into output nodes, removing all edges out of W^o.

Remark 3.2.10. *For a CADMG $G = (J, V, E, L)$, also $G_{\mathrm{swig}(W)}$ is acyclic. If $<$ is any topological order of G given by enumerating all nodes $v \in J \cup V$ via:*

$$v_1 < v_2 < \cdots < v_n,$$

then, for instance, a topological order for $G_{\mathrm{swig}(W)}$ can be achieved by assigning for a node $v_j \in W$ with index j the node v_j^o the index $j - \frac{1}{3}$ and v_j^i the index $j + \frac{1}{3}$, and then ordering all nodes according to their index value.

Lemma 3.2.11 (Two disjoint node-splitting hard interventions commute)**.** *Let $G = (J, V, E, L)$ be a CADMG and $W_1, W_2 \subseteq V$ two disjoint subsets of the output nodes from G. Then the CADMG obtained from first node-splitting on W_1 and then node-splitting* *on W_2 is the same CADMG that arises from first node-splitting on W_2 and then node-splitting on W_1:*

$$\left(G_{\mathrm{swig}(W_1)}\right)_{\mathrm{swig}(W_2)} = \left(G_{\mathrm{swig}(W_2)}\right)_{\mathrm{swig}(W_1)} = G_{\mathrm{swig}(W_1 \dot\cup W_2)}.$$

Lemma 3.2.12 (Disjoint hard interventions and node-splitting hard interventions commute)**.** *Let $G = (J, V, E, L)$ be a CADMG and $W_1 \subseteq J \cup V$ and $W_2 \subseteq V$ two disjoint subsets of nodes from G. Then the CADMG obtained from first hard intervening on W_1 and then node-splitting on W_2 is the same CADMG that arises from first node-splitting on W_2 and then hard intervening on W_1.*

$$\left(G_{\mathrm{do}(W_1)}\right)_{\mathrm{swig}(W_2)} = \left(G_{\mathrm{swig}(W_2)}\right)_{\mathrm{do}(W_1)}.$$

Remark 3.2.13. *Note that if W_1 and W_2 are not disjoint and $w \in W_1 \cap W_2 \subseteq V$ then first hard intervening on w turns w into an input node, for now indicated as w^i, and a node-splitting hard intervention (if we would define it for input nodes) would not change w^i. If, on the other hand, we would first split the node w into w^o and w^i then we would first need to resolve the ambiguity on which of those two the hard intervention should be applied. A hard intervention on w^i would not do anything, but would leave the additional output node w^o in the graph, while hard intervening on w^o would turn w^o into an additional input node, for now indicated as $(w^o)^i$. So in the latter case we are left with two input node $(w^o)^i$, which does not have any edges, and w^i, which might have outgoing edges.*

3.2.4. Intervention Nodes

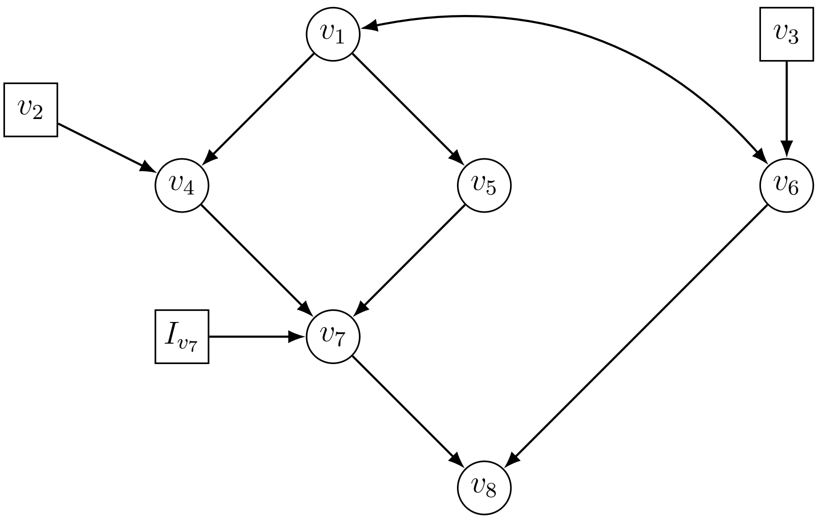


Figure 7: The CADMG from Figure 3 after adding intervention nodes $\{I_{v_2}, I_{v_7}\}$ (where I_{v_2} is identified with v_2 since v_2 is an input node).

More generally, interventions (both hard and soft) can be modeled graphically via auxiliary intervention nodes.

Definition 3.2.14 (Extending CDMGs with intervention nodes)**.** *Let $G = (J, V, E, L)$ be a CDMG and $W \subseteq J \cup V$ a subset of nodes. The **extended CDMG** of G w.r.t. nodes $W \subseteq J \cup V$ and corresponding **intervention nodes** $I_W = \{I_w \,|\, w \in W\}$ with $I_j := j$ for $j \in J \cap W$, is the CDMG:*

$$G_{\mathrm{do}(I_W)} := (J_{\mathrm{do}(I_W)}, V_{\mathrm{do}(I_W)}, E_{\mathrm{do}(I_W)}, L_{\mathrm{do}(I_W)}),$$

where:

i.) $J_{\mathrm{do}(I_W)} := J \,\dot{\cup}\, \{I_w \,|\, w \in W \setminus J\}$,

ii.) $V_{\mathrm{do}(I_W)} := V$,

iii.) $E_{\mathrm{do}(I_W)} := E \,\dot{\cup}\, \{I_w \longrightarrow w \,|\, w \in W \setminus J\}$,

iv.) $L_{\mathrm{do}(I_W)} := L$,

where we just add nodes I_w for $w \in W \setminus J$ and edges $I_w \longrightarrow w$ for $w \in W \setminus J$.

Remark 3.2.15. *If a CDMG $G = (J, V, E, L)$ is acyclic then also $G_{\mathrm{do}(I_W)}$ is acyclic and a topological order for $G_{\mathrm{do}(I_W)}$ is also one for G. Any topological order of G can be extended to one for $G_{\mathrm{do}(I_W)}$, e.g. by putting all the I_w nodes first in the ordering.*

Lemma 3.2.16 (Adding intervention nodes commutes with disjoint hard interventions)**.** *Let $G = (J, V, E, L)$ be a CDMG and $W_1, W_2 \subseteq J \cup V$ two disjoint subsets of nodes from G. Then we have:*

$$\left(G_{\mathrm{do}(I_{W_1})}\right)_{\mathrm{do}(I_{W_2})} = \left(G_{\mathrm{do}(I_{W_2})}\right)_{\mathrm{do}(I_{W_1})} = G_{\mathrm{do}(I_{W_1 \cup W_2})}.$$

We also have:

$$\left(G_{\mathrm{do}(I_{W_1})}\right)_{\mathrm{do}(W_2)} = \left(G_{\mathrm{do}(W_2)}\right)_{\mathrm{do}(I_{W_1})} = G_{\mathrm{do}(I_{W_1}, W_2)}.$$

Lemma 3.2.17 (Adding intervention nodes commutes with disjoint node-splitting hard interventions)**.** *Let $G = (J, V, E, L)$ be a CADMG and $W_1 \subseteq V$ and $W_2 \subseteq J \cup V$ two disjoint subsets of nodes from G. Then the CADMG that arises from first introducing intervention nodes I_{W_2} and then splitting the nodes from W_1 is the same as the CADMG that arises from first splitting the nodes from W_1 and then introducing the intervention nodes I_{W_2}:*

$$\left(G_{\mathrm{swig}(W_1)}\right)_{\mathrm{do}(I_{W_2})} = \left(G_{\mathrm{do}(I_{W_2})}\right)_{\mathrm{swig}(W_1)}.$$

3.2.5. Marginalization of Graphs

Definition 3.2.18 (Marginalization a.k.a. latent projection on CDMGs)**.** *Let $G = (J, V, E, L)$ be a CDMG and $W \subseteq V$ a subset of output nodes. Then the **marginalization of** G **w.r.t.** W or the **latent projection of** G **onto** $J \cup V \setminus W$ is the CDMG:*

$$G^{V \setminus W | J} := G^{\setminus W} := (J^{\setminus W}, V^{\setminus W}, E^{\setminus W}, L^{\setminus W}),$$

where:

i.) $J^{\setminus W} := J$,

ii.) $V^{\setminus W} := V \setminus W$,

iii.) $E^{\setminus W}$ *consists of all directed edges* $\underline{v} \longrightarrow \overline{v}$ *with* $\underline{v}, \overline{v} \in J \cup V \setminus W$ *for which there exists a directed walk in* G*:*

$$\underline{v} \longrightarrow w_1 \longrightarrow \cdots \longrightarrow w_{n-1} \longrightarrow \overline{v},$$

where all intermediate nodes $w_1, \dots, w_{n-1} \in W$ *(if any).*[13]

iv.) $L^{\setminus W}$ *consists of all bi-directed edges* $\underline{v} \leftrightarrow \overline{v}$ *with* $\underline{v}, \overline{v} \in V \setminus W$, $\underline{v} \neq \overline{v}$, *for which there exists a bifurcation in* G*:*

$$\underline{v} \longleftarrow w_1 \longleftarrow \cdots \longleftarrow w_{k-1} \leftrightarrow\!\!* \; w_k \longrightarrow \cdots \longrightarrow w_{n-1} \longrightarrow \overline{v},$$

where all intermediate nodes $w_1, \dots, w_{n-1} \in W$ *(if any).*

Remark 3.2.19. *Marginalization preserves ancestral relations, bifurcations and acyclicity:*

1. *For* $v_1, v_2 \in G$ *with* $v_1, v_2 \notin W$ *we have the equivalence:*

$$v_1 \in \mathrm{Anc}^G(v_2) \qquad \Longleftrightarrow \qquad v_1 \in \mathrm{Anc}^{G^{\setminus W}}(v_2).$$

2. *For* $v_1, v_2 \in G \setminus W$ *(and, optionally,* $v_3 \in G \setminus W$*): there is a bifurcation between* v_1 *and* v_2 *(with source* v_3*) in* G *if and only if there is a bifurcation between* v_1 *and* v_2 *(with source* v_3*) in* $G^{\setminus W}$*.*

3. *If the CDMG* G *is acyclic then so is* $G^{\setminus W}$ *and a topological order of* G *induces a topological order on* $G^{\setminus W}$ *(by just ignoring the nodes from* W*).*

Proof. We prove 2. Let

$$v_1 \longleftarrow w_1 \longleftarrow \cdots \longleftarrow w_{k-1} \leftrightarrow w_k \longrightarrow \cdots \longrightarrow w_{n-1} \longrightarrow w_n \longrightarrow v_2$$

or

$$v_1 \longleftarrow w_1 \longleftarrow \cdots \longleftarrow w_{k-1} \longleftarrow v_3 \longrightarrow w_k \longrightarrow \cdots \longrightarrow w_{n-1} \longrightarrow w_n \longrightarrow v_2$$

be a bifurcation between v_1 and v_2 in G (with source v_3, if applicable). If one marginalizes out a single node $w \in W$ that is not on the bifurcation, then the same bifurcation exists in $G^{\setminus \{w\}}$. If one marginalizes out a single node $w_i \in W$ that appears on the bifurcation one obtains again a bifurcation in $G^{\setminus \{w_i\}}$ (with source v_3, if applicable). By induction it follows that there is a bifurcation between v_1 and v_2 in $G^{\setminus W}$ (with source v_3, if applicable).

For the converse, assume that there exists a bifurcation between v_1 and v_2 in $G^{\setminus W}$ (with source v_3, if applicable). For each directed edge $u \longrightarrow u'$ on this bifurcation, there exists a directed path $u \longrightarrow \dots \longrightarrow u'$ in G with all intermediate nodes in W. Concatenating all these directed paths, one obtains a bifurcation between v_1 and v_2 in $G^{\setminus W}$ (with source v_3, if applicable). ∎

[13] Note that this may introduce self-cycles.

Lemma 3.2.20 (Marginalizations commute)**.** *Let $G = (J, V, E, L)$ be a CDMG and $W_1, W_2 \subseteq V$ two disjoint subsets of output nodes. Then we have:*

$$\left(G^{\setminus W_1}\right)^{\setminus W_2} = \left(G^{\setminus W_2}\right)^{\setminus W_1} = G^{\setminus (W_1 \cup W_2)}.$$

Lemma 3.2.21 (Marginalization and intervention commute)**.** *Let $G = (J, V, E, L)$ be a CDMG and $W_1 \subseteq J \cup V$ and $W_2 \subseteq V$ two disjoint subsets of nodes from G. Then we have:*

$$\left(G_{\mathrm{do}(W_1)}\right)^{\setminus W_2} = \left(G^{\setminus W_2}\right)_{\mathrm{do}(W_1)}.$$

A similar statement holds for marginalizations and adding intervention nodes, and also for marginalizations and node-splitting interventions.

Lemma 3.2.22 (Marginalizing out the output part of splitted nodes equals hard intervention)**.** *Let $G = (J, V, E, L)$ be a CDMG and $W \subseteq V$ be a subset of output nodes from G. Then the CDMG that arises by first splitting the nodes on W and then marginalizing out the nodes from W^o can be identified with the CDMG that arises by hard intervention on W:*

$$G_{\mathrm{do}(W)} \cong \left(G_{\mathrm{swig}(W)}\right)^{\setminus W^o}, \qquad w \mapsto w^i.$$

3.3. σ-Separation

Definition 3.3.1 (Colliders and non-colliders)**.** *Let $G = (J, V, E, L)$ be a CDMG and π a walk in G:*

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

A node v_k, or more precisely, the position $k \in \{0, \dots, n\}$, on the walk π is called:

1. *a **non-collider** on π, if there is at most one arrowhead pointing towards v_k, i.e. if it falls into one of the following cases:*

$$\begin{aligned} \textit{end-node:} &\quad k \in \{0, n\}, \\ \textit{left chain:} &\quad v_{k-1} \leftarrow v_k \leftarrow\!\!\ast\, v_{k+1}, \\ \textit{right chain:} &\quad v_{k-1} \ast\!\!\rightarrow v_k \rightarrow v_{k+1}, \\ \textit{fork:} &\quad v_{k-1} \leftarrow v_k \rightarrow v_{k+1}; \end{aligned}$$

2. *a **collider** on π, if it is of the form:*

$$v_{k-1} \ast\!\!\rightarrow v_k \leftarrow\!\!\ast\, v_{k+1},$$

   i.e. if there are two arrowheads pointing towards v_k on the walk π.

Definition 3.3.2 (Blockable and unblockable non-colliders)**.** *Let $G = (J, V, E, L)$ be a CDMG and π a walk in G:*

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

*We call a non-collider v_k on π an **unblockable non-collider** on π if it is not an end-node ($k \notin \{0, n\}$) and it only has outgoing edges on π to nodes in the same strongly connected component of G. That is, it is one of the following patterns:*

$$\begin{array}{rlll} \textit{left chain:} & v_{k-1} \leftarrow v_k \leftarrow\!\!* \, v_{k+1} & \textit{with} & v_{k-1} \in \mathrm{Sc}^G(v_k) \\ \textit{right chain:} & v_{k-1} *\!\!\rightarrow v_k \rightarrow v_{k+1} & \textit{with} & v_{k+1} \in \mathrm{Sc}^G(v_k) \\ \textit{fork:} & v_{k-1} \leftarrow v_k \rightarrow v_{k+1} & \textit{with} & v_{k-1} \in \mathrm{Sc}^G(v_k) \wedge v_{k+1} \in \mathrm{Sc}^G(v_k) \end{array}$$

*Otherwise, v_k is called a **blockable non-collider** on π. This means that v_k is either an end-node ($k \in \{0, n\}$) or it has at least one outgoing arrow $v_k \rightarrow v_{k\pm1}$ pointing to a node $v_{k\pm1}$ that lies in a different strongly connected component than v_k, i.e. $v_{k\pm1} \notin \mathrm{Sc}^G(v_k)$.*

Remark 3.3.3. *If G is acyclic then all non-colliders are blockable.*

Definition 3.3.4 (σ-blocked walks)**.** *Let $G = (J, V, E, L)$ be a CDMG and $C \subseteq J \cup V$ a subset of nodes and π a walk in G:*

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

We say that the walk π is:

1. *C-σ-**open** (or σ-**open given** C) if and only if:*
   - *i.) all colliders v_k on π are in $\mathrm{Anc}^G(C)$, **and**:*
   - *ii.) all blockable non-colliders v_k on π are **not** in C.*
2. *C-σ-**blocked** (or σ-**blocked given** C) if and only if:*
   - *i.) there exists a collider v_k on π that is **not** in $\mathrm{Anc}^G(C)$, **or**:*
   - *ii.) there exists a blockable non-collider v_k on π in C.*

Note that unblockable non-colliders are always C-σ-open, regardless of the subset $C \subseteq V \cup J$.

Definition 3.3.5 (σ-separation)**.** *Let $G = (J, V, E, L)$ be a CDMG and $A, B, C \subseteq J \cup V$ (not necessarily disjoint) subset of nodes. We then say that:*

1. *A **is σ-separated from** B **given** C **in** G, in symbols:*

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

if every walk from a node in A to a node in $J \cup B$ (sic!)[14] is σ-blocked by C.

[14]The choice to include J here in this place is non-standard in the literature. However, if we include J in this definition here the implied (asymmetric) separoid rules for d-/σ-separation will be of the same form as those for Markov kernels regarding conditional independence. This is the reason we include J here.

2. *If that property does not hold we will write:*

$$A \overset{\sigma}{\underset{G}{\not\perp}} B \,|\, C.$$

3. *We also define the special case:*

$$A \overset{\sigma}{\underset{G}{\perp}} B \qquad :\Longleftrightarrow \qquad A \overset{\sigma}{\underset{G}{\perp}} B \,|\, \emptyset.$$

The following result is often helpful to simplify proofs and to make checking σ-separation feasible in practice.

Proposition 3.3.6. *Let $G = (J, V, E, L)$ be a CDMG. For $C \subseteq J \cup V$, and $w_1, w_2 \in J \cup V$, the following are equivalent:*

1. *there exists a C-σ-open* ***path*** *between w_1 and w_2 in G;*

2. *there exists a C-σ-open* ***walk*** *between w_1 and w_2 in G;*

3. *there exists a C-σ-open* ***walk*** *between w_1 and w_2 in G such that all its colliders lie in C (and not just in $\mathrm{Anc}^G(C)$).*

Remark 3.3.7. 1. *By Proposition 3.3.6 we have that $A \perp^{\sigma}_{G} B \,|\, C$ is equivalent to either of the following:*

   a) *every walk from a node in A to a node in $J \cup B$ is C-σ-blocked by C;*

   b) *every path from a node in A to a node in $J \cup B$ is C-σ-blocked by C.*

2. *Proposition 3.3.6 also shows that if $A \not\perp^{\sigma}_{G} B \,|\, C$ holds then:*

   a) *there exists a (shortest) C-σ-open path from a node in A to a node in $J \cup B$;*

   b) *there exists a (shortest) C-σ-open walk from a node in A to a node in $J \cup B$ such that all its colliders lie in C.*

In practice we usually check if every path is C-σ-blocked or not. This is because there are, in contrast to walks, only a finite number of paths in a (finite) graph. In proofs, though, it often is easier to make use of walks, since these can be concatenated into walks (while one cannot in general concatenate two paths and again obtain a path).

Lemma 3.3.8 (σ-separation under marginalization)**.** *Let $G = (J, V, E, L)$ be a CDMG, $A, B, C \subseteq J \cup V$ and $D \subseteq V$ be subsets of nodes such that:*

$$D \cap (A \cup B \cup C) = \emptyset.$$

Then we have the equivalence:

$$A \overset{\sigma}{\underset{G}{\perp}} B \,|\, C \qquad \Longleftrightarrow \qquad A \overset{\sigma}{\underset{G^{\setminus D}}{\perp}} B \,|\, C.$$

Remark 3.3.9. *If a CDMG G is acyclic then all non-colliders are blockable. So, the partial condition for σ-separation "a blockable non-collider in C" can be simplified to "(any) non-collider in C".*

So in the acyclic case we can simplify the notion of σ-separation, which is usually referred to as ***d-separation****. However, in the non-acyclic setting d-separation ("(any) non-collider in C") and σ-separation ("a blockable non-collider in C") are clearly not equivalent anymore.*

It turned out that in the non-acyclic case σ-separation is the more general concept (and as said above it also captures the acyclic case equivalently well), see [FM17, FM18, FM20, BFPM21]. We will first focus on CADMGs (acyclic) for which we can restrict ourselves to the somewhat simpler d-separation. Later, we will pick up σ-separation again when we deal with cycles.

Proofs - σ-Open Walks and Paths

The following lemma will be convenient to relate σ-open walks and paths in the notion of σ-separation.

Lemma 3.3.10. *Let $G = (J, V, E, L)$ be a CDMG, $C \subseteq V \cup J$ and $\pi = (v_0 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast v_n)$ be a C-σ-open walk in G. Suppose $v_i \in \mathrm{Sc}^G(v_j)$ for some $i, j \in \{0, \dots, n\}$ with $i < j$. If we then replace the subwalk $v_i \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast v_j$ of π by*

(i) a shortest directed path $v_i \longrightarrow \cdots \longrightarrow v_j$ in G if $j = n$ or if $v_j \longrightarrow v_{j+1}$ on π, or

(ii) a shortest directed path $v_i \longleftarrow \cdots \longleftarrow v_j$ in G otherwise,

then this new subwalk is entirely within $\mathrm{Sc}^G(v_j)$ and the modified walk π' is still C-σ-open.

Proof. π' cannot become C-σ-blocked at one of the initial nodes $v_0, \dots, v_{i-1}$ or at one of the final nodes $v_{j+1}, \dots, v_n$ on π', since these nodes occur in the same local configuration on π and are not C-σ-blocked on π by assumption. Furthermore, π' cannot become C-σ-blocked at one of the nodes strictly between v_i and v_j on π' (if there are any), since these nodes are all non-endnode non-colliders that only point to nodes in the same strongly connected component $\mathrm{Sc}^G(v_j)$. It is also worth noting that π' cannot become C-σ-blocked at any of its endnodes, which could be v_i or v_j or both, because those are the same in π. So in the following we can w.l.o.g. assume that both v_i and v_j are non-endnodes of π and thus π'.

Case (i). By assumption v_j is either a fork or a right chain (or the right endnode) on π that is C-σ-open. Since the same blocking criteria apply to v_j on π' it remains C-σ-open on π'. If $v_i = v_j$ then also v_i is C-σ-open on π' (if v_i is the left endnode or not). If $v_i \neq v_j$, then the new directed path $v_i \longrightarrow \cdots \longrightarrow v_j$ in π' is C-σ-open at v_i because all nodes in between lie in the same stronly connected component $\mathrm{Sc}^G(v_i)$ (or v_i is the left endnode anyways).

Case (ii). Since case (i) is solved we can assume that we have $j < n$ with $v_j \leftarrow\!\!\ast v_{j+1}$ in π. If $v_{i-1} \longleftarrow v_i$ on π' (or v_i the left endnode) then this case is analogous to case (i). So we can also assume that we have $i > 0$ and $v_{i-1} \ast\!\!\to v_i$ on π. So π looks as follows:

$$\pi : \qquad \cdots v_{i-1} \ast\!\!\to v_i \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast v_j \leftarrow\!\!\ast v_{j+1} \cdots .$$

So there must be a smallest number $k \in \{i, \dots, j\}$ such that a collider appears at v_k on π:

$$\pi : \qquad \cdots v_{i-1} \ast\!\!\rightarrow v_i \rightarrow \cdots \rightarrow v_k \leftarrow\!\!\ast \cdots \ast\!\!-\!\!\ast\, v_j \leftarrow\!\!\ast\, v_{j+1} \cdots .$$

Since π is C-σ-open we have $v_k \in \mathrm{Anc}^G(C)$. Since $v_i \in \mathrm{Anc}^G(v_k)$ (otherwise v_k would not be the first collider appearing after v_i) we thus have that also $v_i \in \mathrm{Anc}^G(C)$. So if we replace the subwalk $v_i \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast\, v_j$ of π by the shortest directed path $v_i \leftarrow \cdots \leftarrow v_j$ in G we then get for π' the following situation:

$$\pi' : \qquad \cdots v_{i-1} \ast\!\!\rightarrow v_i \leftarrow \cdots \leftarrow v_j \leftarrow\!\!\ast\, v_{j+1} \cdots ,$$

which is then C-σ-open at v_i as $v_i \in \mathrm{Anc}^G(C)$. Note that this holds also when $v_i = v_j$. If $v_i \neq v_j$ then v_j is also C-σ-open on π' as v_j points left to a node in the same strongly connected component as v_j.

So in all cases π' stays C-σ-open. $\square$

Proposition 3.3.6. *Let $G = (J, V, E, L)$ be a CDMG. For $C \subseteq J \cup V$, and $w_1, w_2 \in J \cup V$, the following are equivalent:*

1. *there exists a C-σ-open **path** between w_1 and w_2 in G;*
2. *there exists a C-σ-open **walk** between w_1 and w_2 in G;*
3. *there exists a C-σ-open **walk** between w_1 and w_2 in G such that all its colliders lie in C (and not just in $\mathrm{Anc}^G(C)$).*

Proof. $3 \implies 2$ and $1 \implies 2$ are trivial. Note that paths are walks.

$2 \implies 3$: Suppose there exists a C-σ-open walk π from w_1 to w_2. Then consider a collider $v_{k-1} \ast\!\!\rightarrow v_k \leftarrow\!\!\ast\, v_{k+1}$ on π with $v_k \in \mathrm{Anc}^G(C) \setminus C$. So there exists a non-trivial directed path from v_k to a node $c_k \in C$ with all other nodes not in C. If we then replace the collider at v_k in π by that path and its reverse we get:

$$\cdots \ast\!\!-\!\!\ast\, v_{k-1} \ast\!\!\rightarrow v_k \rightarrow \cdots \rightarrow c_k \leftarrow \cdots \leftarrow v_k \leftarrow\!\!\ast\, v_{k+1} \ast\!\!-\!\!\ast \cdots .$$

This walk is then C-σ-open at all places between v_k on the left and v_k on the right because they are non-colliders not in C. If we do this iteratively for all colliders not in C we get the desired C-σ-open walk where all colliders lie in C.

$2 \implies 1$: Let $\pi = (v_0 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast\, v_n)$ be a C-σ-open walk between nodes $v_0 = w_1$ and $v_n = w_2$ in G. If a node w occurs more than once on π, let v_i be the first node on π and v_j be the last node on π that are in $\mathrm{Sc}^G(w)$. We now use Lemma 3.3.10 to construct a new walk π' from π by replacing the subwalk between v_i and v_j of π by a particular directed path in $\mathrm{Sc}^G(w)$ between v_i and v_j in such a way that π' is still C-σ-open. In π', the number of nodes that occurs more than once is at least one less than in π, and all nodes within $\mathrm{Sc}^G(w)$ occur within a single segment. This replacement procedure can be repeated until no nodes occur more than once. We have then obtained a C-σ-open path between w_1 and w_2. $\square$

3.4. d-Separation

Definition 3.4.1 (d-blocked walks). *Let $G = (J, V, E, L)$ be a CDMG and $C \subseteq J \cup V$ a subset of nodes and π a walk in G:*

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

1. *We say that the walk π is **C-d-blocked** or **d-blocked by C** to emphasize the use of the bi-directed edges.*[15] *if either:*
   - *i.) $v_0 \in C$ or $v_n \in C$ or:*
   - *ii.) there are two adjacent edges in π of one of the following forms:*

$$\begin{array}{rlll} \textit{left chain:} & v_{k-1} \longleftarrow v_k \leftarrow\!\!\ast\, v_{k+1} & \textit{with} & v_k \in C, \\ \textit{right chain:} & v_{k-1} \ast\!\!\rightarrow v_k \longrightarrow v_{k+1} & \textit{with} & v_k \in C, \\ \textit{fork:} & v_{k-1} \longleftarrow v_k \longrightarrow v_{k+1} & \textit{with} & v_k \in C, \\ \textit{collider:} & v_{k-1} \ast\!\!\rightarrow v_k \leftarrow\!\!\ast\, v_{k+1} & \textit{with} & v_k \notin \mathrm{Anc}^G(C). \end{array}$$

2. *We say that the walk π is **C-d-open** if it is not C-d-blocked.*

Remark 3.4.2. *If we consider end-nodes, left chains, right chains and forks as **non-colliders** then we can simply state:*
π is d-blocked by C if and only if it either contains a non-collider in C or a collider not in $\mathrm{Anc}^G(C)$.

Definition 3.4.3 (d-separation). *Let $G = (J, V, E, L)$ be a CDMG and $A, B, C \subseteq J \cup V$ (not necessarily disjoint) subset of nodes. We then say that:*

1. *A **is d-separated from** B **given** C **in** G, in symbols:*

$$A \overset{d}{\underset{G}{\perp}} B \,|\, C,$$

   if every walk from a node in A to a node in $J \cup B$ (sic!)[14] *is C-d-blocked by C.*

2. *If that property does not hold we will write:*

$$A \overset{d}{\underset{G}{\not\perp}} B \,|\, C.$$

Remark 3.4.4. 1. *A similar result from Proposition 3.3.6 holds for d-separation as well.*

2. *d-separation is stable under marginalization, similar to Lemma 3.3.8.*

[15] The "d" here stands for "directional". d-separation was first only used for DAGs (without bi-directed edges). For ADMGs it was then called m-separation in [Ric03] But since the notion of m-separation is arguably the natural extension of d-separation and to avoid introducing more definitions, we will just call it d-separation as well, which will not create any ambiguity.

3.5. Acyclifications

It is possible to reformulate the notion of σ-separation in terms of d-separation on a modified and acyclic graph by making use of the following construction, which will be the main tool to extend the acyclic theory to the cyclic one. The construction was first proposed in the context of CBNs by [Spi94, Spi95].

Definition 3.5.1. *Given a CDMG $G = (J, V, E, L)$, we call a CADMG $G' = (J', V', E', L')$ an* ***acyclification of*** *G if*

(i) G' is acyclic;

(ii) G' has the same input nodes and output nodes as G, i.e. $J' = J$ and $V' = V$;

(iii) for every pair of nodes (i, j) such that $i \notin \mathrm{Sc}^G(j)$:

a) $i \longrightarrow j \in E'$ iff there exists a node $j' \in \mathrm{Sc}^G(j)$ such that $i \longrightarrow j' \in E$;

b) $i \leftrightarrow j \in L'$ iff there exist nodes $i' \in \mathrm{Sc}^G(i), j' \in \mathrm{Sc}^G(j)$ such that $i' \leftrightarrow j' \in L$;

(iv) for every pair of distinct nodes (i, j) such that $i \in \mathrm{Sc}^G(j)$: $i \longrightarrow j \in E'$ or $i \longleftarrow j \in E'$ or $i \leftrightarrow j \in L'$.

The important property of acyclifications is that they can be used to express σ-separation in a (possibly cyclic) graph in terms of d-separation in an acyclification.

Proposition 3.5.2. *Let $G = (J, V, E, L)$ be a CDMG and G' an acyclification of G. Then for $A, B, C \subseteq V \cup J$ (not necessarily disjoint) subsets of nodes we have the equivalence:*

$$A \overset{\sigma}{\underset{G}{\perp}} B \,|\, C \iff A \overset{\sigma}{\underset{G'}{\perp}} B \,|\, C \iff A \overset{d}{\underset{G'}{\perp}} B \,|\, C.$$

Proof. We will show that there is a C-σ-open walk between A and $B \cup J$ in G if and only if there is a C-σ-open walk between A and $B \cup J$ in G'. Since G' is acyclic, this is in turn equivalent to the existence of a C-d-open walk between A and $B \cup J$ in G'.

$\Longrightarrow$: Suppose there is a C-σ-open walk $\pi = (v_0, \dots, v_n)$ between A and $B \cup J$ in G. All its colliders are in C and all its non-colliders are either not in C, or otherwise, point only to nodes in the same strongly connected component. Note that each edge between two nodes in different strongly connected components in G is also present in G'. Edges between two nodes in the same strongly connected component, however, may not be present in G'. Therefore, we will replace these edges with walks in G'. Consider a subwalk $(v_i, \dots, v_j)$ of maximum length that is entirely contained within a strongly connected component in G (with possibly $i = j$). We distinguish different cases and show for each case how this subwalk can be replaced by a subwalk in G'.

(i) $\ast\!\!\rightarrow v_i \cdots v_j \leftarrow\!\!\ast$: the subwalk between v_i and v_j has to contain a collider, say w, which must be in C since the walk between v_i and v_j is C-σ-open. We can replace this subwalk by $\ast\!\!\rightarrow w \leftarrow\!\!\ast$ in G' such that w becomes a collider in C.

(ii) $(\leftarrow)v_i \cdots v_j \leftarrow\!\!*$:[16] here v_i is a non-collider pointing to another strongly connected component or v_i is an endnode, and in both cases, $v_i \notin C$. Therefore, we can replace the subwalk by $(\leftarrow)v_i \leftarrow\!\!*$ in G', such that v_i becomes a non-collider not in C.

(iii) $*\!\!\rightarrow v_i \cdots v_j(\rightarrow)$: analogous to the previous case, we can replace it by $*\!\!\rightarrow v_j(\rightarrow)$ in G', such that v_j becomes a non-collider not in C.

(iv) $(\leftarrow)v_i \cdots v_j(\rightarrow)$: v_i, v_j are both not in C by assumption. If $i = j$, we replace this subwalk by $(\leftarrow)v_i(\rightarrow)$ such that v_i becomes a non-collider not in C. If $i < j$, we replace this subwalk by $(\leftarrow)v_i *\!\!-\!\!* v_j(\rightarrow)$ with $v_i *\!\!-\!\!* v_j$ any edge connecting v_i and v_j in G', such that both v_i and v_j become non-colliders not in C.

By replacing all maximal subwalks of the original walk π that are contained within a single strongly connected component of G in this way, we obtain a walk in the acyclification G' that is C-σ-open by construction. Note that the modified walk has the same endpoints (v_0 and v_n) as the original walk.

$\Longleftarrow$: Suppose there is a C-σ-open walk π' between A and $B \cup J$ in G'. All its colliders are in C, and all its non-colliders are not in C. We will construct a walk π in G with the same endpoints as π' that is C-σ-open.

Consider a non-trivial subwalk $(v_i, \ldots, v_j)$ on π' of maximum length that is entirely contained within a strongly connected component of G. This subwalk may not be present in G'. We distinguish different cases and show for each case how this subwalk can be replaced by a subwalk in G.

(i) $*\!\!\rightarrow v_i \cdots v_j \leftarrow\!\!*$: the subwalk between v_i and v_j has to contain a collider, say w, which must be in C since the walk between v_i and v_j is C-σ-open, and must be in $\mathrm{Sc}^G(v_i) = \mathrm{Sc}^G(v_j)$ by assumption. We can replace this subwalk by $*\!\!\rightarrow v_i \rightarrow \ldots \rightarrow w \leftarrow \ldots \leftarrow v_j \leftarrow\!\!*$ in G, with possibly $v_i = w$ and possibly $w = v_j$, with all nodes in $\mathrm{Sc}^G(v_i)$. Note that the modified walk remains C-σ-open.

(ii) $(\leftarrow)v_i \cdots v_j \leftarrow\!\!*$: here v_i is a non-collider pointing to another strongly connected component or v_i is an endnode, and in both cases, $v_i \notin C$. We can replace this subwalk by a shortest directed walk $(\leftarrow)v_i \leftarrow \ldots \leftarrow v_j \leftarrow\!\!*$ in G with all nodes in $\mathrm{Sc}^G(v_i)$. Note that the modified walk remains C-σ-open.

(iii) $*\!\!\rightarrow v_i \cdots v_j(\rightarrow)$: analogous to the previous case, we can replace it by $*\!\!\rightarrow v_i \rightarrow \ldots \rightarrow v_j(\rightarrow)$ in G.

(iv) $(\leftarrow)v_i \cdots v_j(\rightarrow)$: v_i, v_j are both not in C by assumption. We can replace this subwalk by a shortest directed walk $(\leftarrow)v_i \rightarrow \ldots \rightarrow v_j(\rightarrow)$ in G with all nodes in $\mathrm{Sc}^G(v_i)$. The modified walk remains C-σ-open.

[16] We put parentheses around the first directed edge to indicate that this case also applies if v_i is an endnode, i.e., if $i = 0$.

In each of the four cases, in the modified walk both v_i and v_j become either colliders in C, or non-colliders not in C, or non-colliders in C that only point to a node in the same strongly connected component of G.

Now, we will replace edges on π' between two strongly connected components that are not present in G. For any directed edge $i \to j$ on π' with $j \notin \mathrm{Sc}^G(i)$, there must be a $j' \in \mathrm{Sc}^G(j)$ such that $i \to j'$ is present in G, and hence there must be a directed path $j' \to \ldots \to j$ entirely in $\mathrm{Sc}^G(j)$ such that we can use $i \to j' \to \ldots \to j$ as replacement in G of the edge $i \to j$. Similarly, for any bidirected edge $i \leftrightarrow j$ on π' with $j \notin \mathrm{Sc}^G(i)$, there must be $i' \in \mathrm{Sc}^G(i)$ and $j' \in \mathrm{Sc}^G(j)$ such that $i' \leftrightarrow j'$ is present in G, and hence there must be a walk $i \leftarrow \ldots \leftarrow i' \leftrightarrow j' \to \ldots \to j$ in G, where $i \leftarrow \ldots \leftarrow i'$ is entirely in $\mathrm{Sc}^G(i)$ and $j' \to \ldots \to j$ is entirely in $\mathrm{Sc}^G(j)$, that we can use as replacement in G of the edge $i \leftrightarrow j$. The new nodes introduced on π in these replacements are non-colliders that only point to nodes in the same strongly connected component. The endpoints of the replacement paths do not change their status: if they were colliders in C on π' they still are on π, and if they were non-colliders not in C on π' they still are on π.

Hence we have constructed a walk π in G with the same endpoints as π' that is C-σ-open. □

The following construction shows that acyclifications exist (but it is just one out of many possible ways to construct acyclifications).

Example 3.5.3 (The standard acyclification)**.** *Let $G = (J, V, E, L)$ be a CDMG. Then we define the* ***standard acyclification*** *of G as the CDMG $G' = (J, V, E', L')$ where:*

$$\begin{aligned}
E' &:= \left\{v_1 \to v_2 \,\middle|\, v_1 \in J \cup V, v_2 \in V, v_2 \notin \mathrm{Sc}^G(v_1), \exists v_2' \in \mathrm{Sc}^G(v_2) :\, v_1 \to v_2' \in E\right\}, \\
L' &:= \left\{v_1 \leftrightarrow v_2 \,\middle|\, v_1, v_2 \in V, v_1 \neq v_2, \exists v_1' \in \mathrm{Sc}^G(v_1), v_2' \in \mathrm{Sc}^G(v_2) :\, v_1 \leftrightarrow v_2 \in L\right\} \\
&\quad \cup \left\{v_1 \leftrightarrow v_2 \,\middle|\, v_1, v_2 \in V, v_1 \neq v_2, v_1 \in \mathrm{Sc}^G(v_2)\right\}.
\end{aligned}$$

The standard acyclification of a CDMG is acyclic, i.e. a CADMG.

Proof. Assume that G' is not acyclic. Then there exists a non-trivial cyclic directed walk in G':

$$v_1 \to v_2 \to \cdots \to v_k \to v_1,$$

for some $k \geq 1$. It is clear that $k \geq 2$ because clearly $v_1 \in \mathrm{Sc}^G(v_1)$, which rules out the existence of an edge $v_1 \to v_1 \in G'$. For simplicity we now identify $v_{k+1} := v_1$ in the following. By construction of G' for every $i = 1, \ldots, k$ there exists $v_{i+1}' \in \mathrm{Sc}^G(v_{i+1})$ such that the edge $v_i \to v_{i+1}'$ exists in G. Since $v_{i+1}' \in \mathrm{Sc}^G(v_{i+1})$ there also exists a directed walk in G:

$$v_{i+1}' \to \cdots \to v_{i+1}.$$

Concatenating all directed walks we get the cyclic directed walk in G:

$$v_1 \to v_2' \to \cdots \to v_2 \to v_3' \to \cdots \to v_k \to v_1' \to \cdots \to v_1.$$

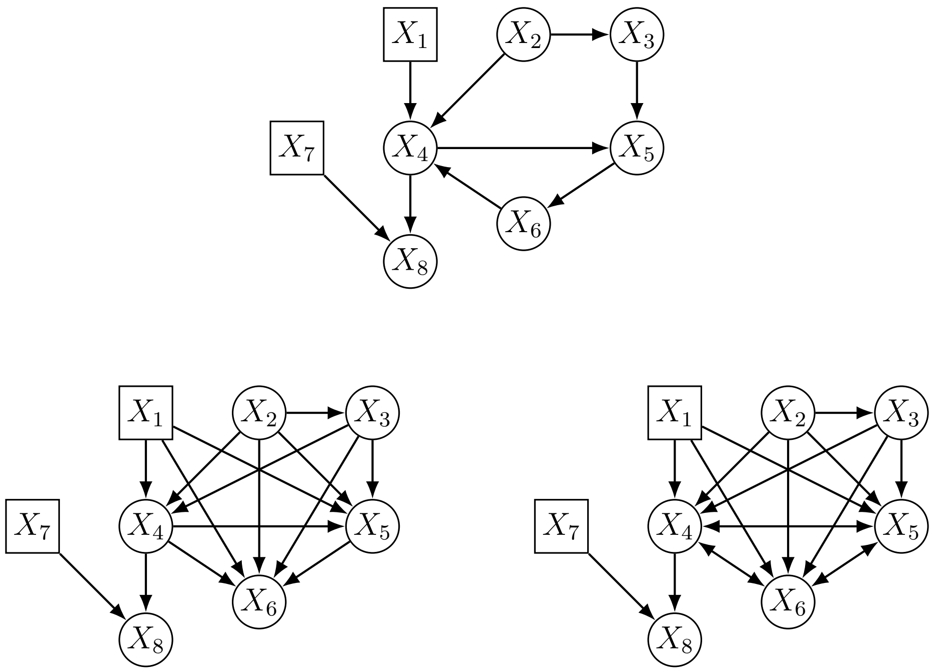


Figure 8: Top: CDMG G. Bottom: two acyclifications of G.

This shows that $v_2 \in \mathrm{Sc}^G(v_1)$, which is a contradiction to the existence of the edge $v_1 \longrightarrow v_2 \in G'$. So a non-trivial cyclic directed walk in G' cannot exist in the first place. So G' must be acyclic. $\square$

3.6. Separoid Axioms for σ-/d-Separation

Definition/Theorem 3.6.1 ((Asymmetric) separoid axioms for σ-separation/d-separation)**.** *Let $G = (J, V, E, L)$ be a CDMG and $A, B, C, D \subseteq J \cup V$ subsets of nodes. Then the ternary relations $\perp = \perp^d_G$ and $\perp = \perp^\sigma_G$ satisfy the following rules:*

a) Extended Left Redundancy:

$$D \subseteq A \implies D \perp B \mid A.$$

b) J-Restricted Right Redundancy:

$A \perp \emptyset \mid C \cup J$ *always holds.*

c) J-Inverted Right Decomposition:

$$A \perp B \mid C \implies A \perp J \cup B \mid C.$$

d) Left Decomposition:

$$A \cup D \perp B \mid C \implies D \perp B \mid C.$$

e) Right Decomposition:

$$A \perp B \cup D \,|\, C \implies A \perp D \,|\, C.$$

f) Left Weak Union:

$$A \cup D \perp B \,|\, C \implies A \perp B \,|\, D \cup C.$$

g) Right Weak Union:

$$A \perp B \cup D \,|\, C \implies A \perp B \,|\, D \cup C.$$

h) Left Contraction:

$$(A \perp B \,|\, D \cup C) \wedge (D \perp B \,|\, C) \implies A \cup D \perp B \,|\, C.$$

i) Right Contraction:

$$(A \perp B \,|\, D \cup C) \wedge (A \perp D \,|\, C) \implies A \perp B \cup D \,|\, C.$$

j) Right Cross Contraction:

$$(A \perp B \,|\, D \cup C) \wedge (D \perp A \,|\, C) \implies A \perp B \cup D \,|\, C.$$

k) Flipped Left Cross Contraction:

$$(A \perp B \,|\, D \cup C) \wedge (B \perp D \,|\, C) \implies B \perp A \cup D \,|\, C.$$

In particular, we have the equivalences:

$$\begin{aligned} (A \perp B \cup D \,|\, C) &\iff (A \perp B \,|\, D \cup C) \quad \wedge \quad (A \perp D \,|\, C), \\ (A \cup D \perp B \,|\, C) &\iff (A \perp B \,|\, D \cup C) \quad \wedge \quad (D \perp B \,|\, C). \end{aligned}$$

We also get:

l) J-Restricted Symmetry:

$$A \perp B \,|\, C \cup J \implies B \perp A \,|\, C \cup J.$$

For the special case of $J = \emptyset$ we have thus (unrestricted) Symmetry.

Remark 3.6.2. *Let the assumptions be like in Theorem 3.6.1. We also have the following rules:*

m) Left Composition:

$$(A \perp B \,|\, C) \wedge (D \perp B \,|\, C) \implies A \cup D \perp B \,|\, C.$$

n) Right Composition:

$$(A \perp B \,|\, C) \wedge (A \perp D \,|\, C) \implies A \perp B \cup D \,|\, C.$$

o) Left Intersection: If $A \cap D = \emptyset$ then:

$$(A \perp B \,|\, D \cup C) \wedge (D \perp B \,|\, A \cup C) \implies A \cup D \perp B \,|\, C.$$

p) Right Intersection: If $B \cap D = \emptyset$ then:

$$(A \perp B \,|\, D \cup C) \wedge (A \perp D \,|\, B \cup C) \implies A \perp B \cup D \,|\, C.$$

Proofs - Separoid Axioms for σ-/d-Separation

In the following let $G = (J, V, E, L)$ be a CDMG and $A, B, C, D \subseteq J \cup V$ (not necessarily disjoint) subsets of nodes.

Recall that we say that A *is σ-separated from B given C* in G, in symbols:

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

if every walk from a node in A to a node in $J \cup B$ (sic!) is σ-blocked by C.
Again, a walk π is *σ-blocked by C* if it either contains a blockable non-collider in C or a collider not in C.

We abbreviate the ternary relations in the following as: $\perp := \overset{\sigma}{\underset{G}{\perp}}$.

Lemma 3.6.3 (Extended Left Redundancy)**.**

$$D \subseteq A \implies D \perp B \,|\, A.$$

Proof. If π is a walk from a node v in D to a node w in $J \cup B$ then its first end node is in A, so π is σ-blocked by A. □

Lemma 3.6.4 (J-Restricted Right Redundancy)**.**

$$A \perp \emptyset \,|\, C \cup J \qquad \textit{always holds.}$$

Proof. If π is a walk from a node v in A to a node w in J then its last end node is in $C \cup J$, so π is σ-blocked by $C \cup J$. □

Lemma 3.6.5 (J-Inverted Right Decomposition)**.**

$$A \perp B \,|\, C \implies A \perp J \cup B \,|\, C.$$

Proof. If π is a walk from a node v in A to a node w in $J \cup J \cup B$ then $w \in J \cup B$. If $w \in J \cup B$ then by assumption π is σ-blocked by C. □

Lemma 3.6.6 (Left Decomposition)**.**

$$A \cup D \perp B \,|\, C \implies D \perp B \,|\, C.$$

Proof. If π is a walk from a node v in D to a node w in $J \cup B$, then the walk π is also a walk from $A \cup D$ to $J \cup B$, which by assumption is σ-blocked by C. □

Lemma 3.6.7 (Right Decomposition)**.**

$$A \perp B \cup D \,|\, C \implies A \perp D \,|\, C.$$

Proof. If π is a walk from a node v in A to a node w in $J \cup D$, then the walk π is also a walk from A to $J \cup B \cup D$, which by assumption is σ-blocked by C. $\square$

Lemma 3.6.8 (Left Weak Union)**.**

$$A \cup D \perp B \,|\, C \implies A \perp B \,|\, D \cup C.$$

Proof. Let us assume the contrary: $A \not\perp B \,|\, D \cup C$. Then there exists a shortest $(D \cup C)$-σ-open walk π from a node v in A to a node w in $J \cup B$ in G such that every collider of π is in $D \cup C$. Then every blockable non-collider of π is not in $D \cup C$.
If now π does not contain any node from $D \setminus C$ then every collider of π is in C. This implies that π is C-σ-open, which contradicts the assumption: $A \cup D \perp B \,|\, C$.
So we can assume now that π contains a node in $D \setminus C$. Then consider the shortest sub-walk $\tilde{\pi}$ in π starting from the end-node $w \in J \cup B$ and going back to the first node $u \in D \setminus C$. This means that $\tilde{\pi}$ is a walk from $D \setminus C$ to $J \cup B$ where the end-node u of $\tilde{\pi}$ is the only node in $D \setminus C$. So $\tilde{\pi}$ does not contain any collider in $D \setminus C$. So all colliders of $\tilde{\pi}$ lie in C. All blockable non-colliders of $\tilde{\pi}$ that are different from the end-node u are also blockable non-colliders on π. They are thus not in $D \cup C$ by the assumption on π, in particular, not in C. The only remaining blockable non-collider u of $\tilde{\pi}$ lies in $D \setminus C$ by construction and it thus lies not in C either. So $\tilde{\pi}$ is C-σ-open walk from $A \cup D$ to $J \cup B$. This contradicts the assumption: $A \cup D \perp B \,|\, C$.
So the premise: $A \not\perp B \,|\, D \cup C$, must be false. This shows: $A \perp B \,|\, D \cup C$. $\square$

Lemma 3.6.9 (Right Weak Union)**.**

$$A \perp B \cup D \,|\, C \implies A \perp B \,|\, D \cup C.$$

Proof. Follow the same steps as in Left Weak Union (Lemma 3.6.8). Soe there exists a shortest $(D \cup C)$-σ-open walk π from a node v in A to a node w in $J \cup B \cup D$ in G such that every collider of π is in $D \cup C$. If π does not contain any nodes from $D \setminus C$ we get a contradiction to: $A \perp B \cup D \,|\, C$. Then, again, we can assume that π contains a node in $D \setminus C$. Then consider the shortest sub-walk $\tilde{\pi}$ in π from $v \in A$ to a node $u \in D \setminus C$. This means that $\tilde{\pi}$ does not contain any collider in $D \setminus C$, so they are all in C. Furthermore, all blockable non-colliders are not in C. So $\tilde{\pi}$ is C-σ-open walk from A to $J \cup B \cup D$. This contradicts the assumption: $A \perp B \cup D \,|\, C$. $\square$

Lemma 3.6.10 (Left Contraction)**.**

$$(A \perp B \,|\, D \cup C) \wedge (D \perp B \,|\, C) \implies A \cup D \perp B \,|\, C.$$

Proof. Let us assume the contrary: $A \cup D \not\perp B \,|\, C$. Then there exists a shortest C-σ-open walk π from a node v in $A \cup D$ to a node w in $J \cup B$ in G such that every collider of π is in C. So every blockable non-collider is not in C. In particular, $v \notin C$. We now

claim that v is the only node of π that is in $(A \cup D) \setminus C$. Otherwise, there would be a non-end-node u of π with $u \in (A \cup D) \setminus C$. Since $u \notin C$ the whole sub-walk from u to w would already be a C-σ-open walk from $A \cup D$ to $J \cup B$, which is also shorter than π, which contradicts the assumption. So we can assume that v is the only node of π that is in $(A \cup D) \setminus C$. In particular, all blockable non-colliders of π that are different from v are not in $D \setminus C$ and thus are not in $D \cup C = (D \setminus C) \cup C$.
Furthermore, v cannot lie in $D \setminus C$ as it would contradict the assumption: $D \perp B \,|\, C$. It follows that $v \in A \setminus C$ and π is a walk from A to $J \cup B$ whose colliders are in $C \subseteq D \cup C$ and all blockable non-colliders are not in $D \cup C$. But this contradicts the other assumption: $A \perp B \,|\, D \cup C$. □

Lemma 3.6.11 (Right Contraction)**.**

$$(A \perp B \,|\, D \cup C) \wedge (A \perp D \,|\, C) \implies A \perp B \cup D \,|\, C.$$

Proof. Let us assume the contrary: $A \not\perp B \cup D \,|\, C$. Then there exists a shortest C-σ-open walk π from a node v in A to a node w in $J \cup B \cup D$ in G such that every collider of π is in C. So every blockable non-collider is not in C and w is the only node of π that is in $(J \cup B \cup D) \setminus C$ (otherwise π could be shortened).
Also w cannot lie in $D \setminus C$ as it would contradict the assumption: $A \perp D \,|\, C$. Thus $w \in (J \cup B) \setminus C$ and π is a walk from A to $J \cup B$ whose colliders all are in $C \subseteq D \cup C$ and all blockable non-colliders are not in $D \cup C$. But this contradicts the other assumption: $A \perp B \,|\, D \cup C$. □

Lemma 3.6.12 (Right Cross Contraction)**.**

$$(A \perp B \,|\, D \cup C) \wedge (D \perp A \,|\, C) \implies A \perp B \cup D \,|\, C.$$

Proof. Verbatim the same as Right Contraction (Lemma 3.6.11), only the first contradiction is with: $D \perp A \,|\, C$. □

Lemma 3.6.13 (Flipped Left Cross Contraction)**.**

$$(A \perp B \,|\, D \cup C) \wedge (B \perp D \,|\, C) \implies B \perp A \cup D \,|\, C.$$

Proof. Let us assume the contrary: $B \not\perp A \cup D \,|\, C$. Then there exists a shortest C-σ-open walk π from a node v in B to a node w in $J \cup A \cup D$ in G such that every collider of π is in C. So every blockable non-collider is not in C and w is the only node of π that is in $(J \cup A \cup D) \setminus C$ (otherwise π could be shortened).
Also w cannot lie in $(J \cup D) \setminus C$ as it would contradict the assumption: $B \perp D \,|\, C$. Thus $w \in A \setminus C$ and the walk π (in reverse direction) is a walk from A to B whose colliders are all in $C \subseteq D \cup C$ and all blockable non-colliders are not in $D \cup C$. But this contradicts the other assumption: $A \perp B \,|\, D \cup C$. □

Lemma 3.6.14 (J-Restricted Symmetry)**.**

$$A \perp B \,|\, C \cup J \implies B \perp A \,|\, C \cup J.$$

Proof. This follows from Flipped Left Cross Contraction (Lemma 3.6.13) with $D = \emptyset$ and $C \cup J$ in place of C together with J-Restricted Right Redundancy (Lemma 3.6.4). $\square$

Lemma 3.6.15 (Left Composition)**.**

$$(A \perp B \,|\, C) \wedge (D \perp B \,|\, C) \implies A \cup D \perp B \,|\, C.$$

Proof. Let π be a walk from a node v in $A \cup D$ to a node w in $J \cup B$. If $v \in A$ then π is σ-blocked by C by assumption: $A \perp B \,|\, C$. If $v \in D$ then π is σ-blocked by C by assumption: $D \perp B \,|\, C$. $\square$

Lemma 3.6.16 (Right Composition)**.**

$$(A \perp B \,|\, C) \wedge (A \perp D \,|\, C) \implies A \perp B \cup D \,|\, C.$$

Proof. Let π be a walk from a node v in A to a node w in $J \cup B \cup D$. If $w \in J \cup B$ then π is σ-blocked by C by assumption: $A \perp B \,|\, C$. If $w \in J \cup D$ then π is σ-blocked by C by assumption: $A \perp D \,|\, C$. $\square$

Lemma 3.6.17 (Left Intersection)**.** *Assume that $A \cap D = \emptyset$, then:*

$$(A \perp B \,|\, D \cup C) \wedge (D \perp B \,|\, A \cup C) \implies A \cup D \perp B \,|\, C.$$

Proof. Let us assume the contrary: $A \cup D \not\perp B \,|\, C$. Then there exists a shortest C-σ-open walk π from a node v in $A \cup D$ to a node w in $J \cup B$ in G such that every collider of π is in C. So every blockable non-collider is not in C and v is the only node of π that is in $(A \cup D) \setminus C$ (otherwise π could be shortened).
If $v \in A$ then by the disjointness of A and D we have that $v \notin D$. Then π is a walk from A to $J \cup B$ whose colliders are in $C \subseteq D \cup C$ and all blockable non-colliders are not in $(D \setminus C) \cup C = D \cup C$. This contradicts the assumption: $A \perp B \,|\, D \cup C$.
If $v \in D$ then similarly we get a contradiction: $D \perp B \,|\, A \cup C$ $\square$

Lemma 3.6.18 (Right Intersection)**.** *Assume that $B \cap D = \emptyset$, then:*

$$(A \perp B \,|\, D \cup C) \wedge (A \perp D \,|\, B \cup C) \implies A \perp B \cup D \,|\, C.$$

Proof. Let us assume the contrary: $A \not\perp B \cup D \,|\, C$. Then there exists a shortest C-σ-open walk π from a node v in A to a node w in $J \cup B \cup D$ in G such that every collider of π is in C. So every blockable non-collider is not in C and w is the only node of π that is in $(J \cup B \cup D) \setminus C$ (otherwise π could be shortened).
If $w \notin B$ then $w \in J \cup D$. In this case π is a walk from A to $J \cup D$ where every collider is in $C \subseteq B \cup C$ and all blockable non-colliders are not in $(B \setminus C) \cup C = B \cup C$. So π is a $(B \cup C)$-σ-open walk from A to $J \cup D$. This contradicts the assumption: $A \perp D \,|\, B \cup C$.
If $w \notin D$ then $w \in J \cup B$. In this case π is a walk from A to $J \cup B$ where every collider is in $C \subseteq D \cup C$ and all blockable non-colliders are not in $D \cup C$. So π is a $(D \cup C)$-σ-open walk from A to $J \cup B$. This contradicts the assumption: $A \perp B \,|\, D \cup C$.
Since $B \cap D = \emptyset$ there are no other cases ($B^{\mathsf{c}} \cup D^{\mathsf{c}} = J \cup V$) and we are done. $\square$

Remark 3.6.19 (Proofs for the separoid axioms for d-separation)**.** *The proofs for the separoid axioms for d-separation are verbatim the same as above if one exchanges the word "blockable non-collider" with just the word "non-collider" everywhere.*

4. Causal Bayesian Networks

4.1. Core Concepts

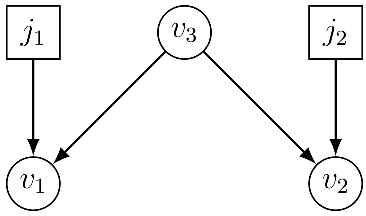


Figure 9: The Conditional Directed Acyclic Graph (CDAG) of a Causal Bayesian Network (CBN) with input variables.

Definition 4.1.1 (Causal Bayesian network)**.** *A* ***causal Bayesian network (CBN)****—by definition—consists of:*

a) a conditional directed acyclic graph (CDAG): $G = (J, V, E)$ (with finite vertex sets, and no bidirected edges),

b) a standard measurable space $\mathcal{X}_v$ for every $v \in J \cup V$,

c) for every $v \in V$, a Markov kernel: $P_v\left(X_v | X_{\mathrm{Pa}^G(v)}\right)$:

$$\begin{array}{rcl} \mathcal{X}_{\mathrm{Pa}^G(v)} & \dashrightarrow & \mathcal{X}_v, \\ (A, x_{\mathrm{Pa}^G(v)}) & \mapsto & P_v\left(X_v \in A | X_{\mathrm{Pa}^G(v)} = x_{\mathrm{Pa}^G(v)}\right), \end{array}$$

where we write for $D \subseteq J \cup V$:

$$\begin{aligned} \mathcal{X}_D &:= \prod_{v \in D} \mathcal{X}_v, & \mathcal{X}_\emptyset &:= * = \{*\}, \\ X_D &:= (X_v)_{v \in D}, & X_\emptyset &:= *, \\ x_D &:= (x_v)_{v \in D}, & x_\emptyset &:= *. \end{aligned}$$

Remark 4.1.2. *Most existing accounts of causal Bayesian networks do not formally distinguish input nodes from output nodes. The reasons that we do make this disctinction are of a measure-theoretical nature. If all variables are discrete, and all probability mass functions and Markov kernels are strictly positive, then the formal differences between input and output nodes may be ignored and everything can be considered as output nodes.*

Definition 4.1.3 (The joint Markov kernel of a causal Bayesian network with input variables)**.** *Consider a causal Bayesian network with input variables with CDAG $G = (J, V, E)$ with Markov kernels $P_v\left(X_v | X_{\mathrm{Pa}^G(v)}\right)$ for $v \in V$. For a fixed topological ordering $<$ of G we then define the joint Markov kernel of the CBN:*

$$\mathcal{X}_J \dashrightarrow \mathcal{X}_V$$

as follows:

$$P(X_V|\,\mathrm{do}(X_J)) := \bigotimes_{v\in V}^{>} P_v\left(X_v|X_{\mathrm{Pa}^G(v)}\right),$$

where the nodes v run through V in reverse ordering of $<$, i.e. all parents are on the right of all their children.

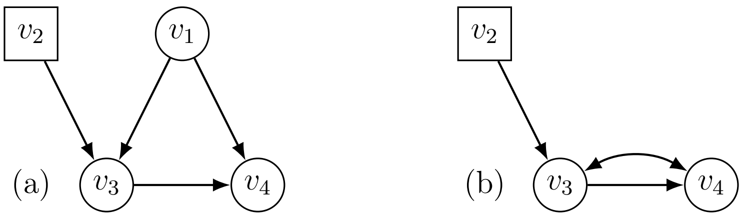


Figure 10: (a) Conditional Directed Acyclic Graph (CDAG) G; (b) Conditional Acyclic Directed Mixed Graph (CADMG) $G^{\setminus\{v_1\}}$ obtained after marginalizing out v_1.

Example 4.1.4. *The CDAG G displayed in Figure 10(a) and Markov kernels $P_1(X_1)$, $P_3(X_3|X_1,X_2)$, $P_4(X_4|X_1,X_3)$ give a joint Markov kernel of a CBN:*

$$P(X_1,X_3,X_4|\,\mathrm{do}(X_2)) = P_4(X_4|X_1,X_3)\otimes P_3(X_3|X_1,X_2)\otimes P_1(X_1).$$

Exercise 4.1.5. *Show that the definition of the joint Markov kernel of a CBN is independent of the topological ordering.*

Notation 4.1.6. *By abuse of notation, we will refer to the tuple:*

$$M = \Big(G=(J,V,E), \big(P_v(X_v|X_{\mathrm{Pa}^G(v)})\big)_{v\in V}\Big),$$

or just to the tuple:

$$M = (G, P\left(X_V|\,\mathrm{do}(X_J)\right))$$

as the CBN, keeping the single Markov kernels $P_v(X_v|X_{\mathrm{Pa}^G(v)})$ and the spaces $\mathcal{X}_v$ implicit.

Remark 4.1.7 (Marginalization and conditioning)**.** *Let $P(X_V|\,\mathrm{do}(X_J))$ be the joint Markov kernel of a CBN. We can extend it to a joint Markov kernel including X_J:*

$$P(X_V,X_J|\,\mathrm{do}(X_J)) = P(X_V|\,\mathrm{do}(X_J))\otimes\delta(X_J|X_J).$$

For any $A,B\subseteq J\cup V$ we then also have the ***marginal conditional*** *Markov kernel:*

$$P(X_A|X_B,\mathrm{do}(X_J)),$$

which exists by theorem 2.4.16 due to the use of standard measurable spaces and is unique up to a $P(X_B|\,\mathrm{do}(X_J))$-null set.
Furthermore, if $C\subseteq J$ and we have:

$$X_A \underset{P(X_V|\,\mathrm{do}(X_J))}{\perp\!\!\!\perp} X_J\,|\,X_C,$$

then we also have a Markov kernel:

$$P(X_A|\,\mathrm{do}(X_C))$$

that fits into the equation:

$$P(X_A, X_C|\,\mathrm{do}(X_J)) = P(X_A|\,\mathrm{do}(X_C)) \otimes P(X_C|\,\mathrm{do}(X_J)).$$

Note that this $P(X_A|\,\mathrm{do}(X_C))$ is unique up to a $P(X_C|\,\mathrm{do}(X_J))$-null set. Since, further, $P(X_C|\,\mathrm{do}(X_J)) = \delta(X_C|X_J)$, we even get that $P(X_A|\,\mathrm{do}(X_C))$ is unique (not just up to null sets). In other words, the above conditional independence states that $P(X_A|\,\mathrm{do}(X_J))$ is only dependent on the arguments from X_C and can be represented by a Markov kernel $P(X_A|\,\mathrm{do}(X_C))$.

Definition 4.1.8 (Causal Bayesian network with latent variables)**.** *A **causal Bayesian network with latent variables (L-CBN)**— per definition—consists of a CBN:*

$$M = \left(G^+ = (J, V^+, E^+), \left(P_v(X_v|X_{\mathrm{Pa}^{G^+}(v)})\right)_{v \in V^+}\right),$$

*together with a disjoint decomposition of the output nodes $V^+ = V \dot{\cup} U$ into **observed nodes** V and **unobserved nodes** U.*

Remark 4.1.9. *In the Definition 4.1.8 we make the distinction between the set of observed nodes V and that of unobserved nodes U. We could have made that distinction already earlier in the graph theory chapters and introduce CDAGs $G^+ = (J, (V, U), E^+)$, where we make the distinction between these node types part of the (or a new) definition. However, most of the time these sets are mathematically treated the same way and we could just consider their union $V^+ = V \dot{\cup} U$. Usually the distinction between V and U is only made to indicate which variables are marginalized out. Also, it often happens that one considers the same CBN in different settings, and which variables are observed and unobserved depends on the setting (for example, during training of a classifier both features and labels are observed, while during testing only features are observed). For all these reasons, we do not consider the specification of which variables are observed and which are latent part of the model.*

Notation 4.1.10.

1. *We will also often just denote a causal Bayesian network with latent variables by the tuple:*

$$M = \left(G^+ = (J, (V, U), E^+), \left(P_v(X_v|X_{\mathrm{Pa}^{G^+}(v)})\right)_{v \in V \cup U}\right),$$

or just:

$$M = \left(G^+,\ P\left(X_{V \cup U}|\,\mathrm{do}\left(X_J\right)\right)\right).$$

2. *We refer to the marginal Markov kernel of M:*

$$P\left(X_V|\,\mathrm{do}\left(X_J\right)\right)$$

*as the **observable Markov kernel**.*

3. *We call the marginalized CADMG of M:*

$$G := (J, V, E, L) := (G^+)^{\setminus U}$$

the (induced) ***observable CADMG****.*

4. *We will often just refer to M as "a CBN with observed nodes V" or "a CBN with latent nodes U" or "a CBN with observed CADMG G" to mean that M is a causal Bayesian network with latent variables with latent nodes U and observed nodes V.*

Example 4.1.11. *Consider again the CADG G displayed in Figure 10(a) (see also Example 4.1.4). If we assume v_1 to be a latent variable (and v_3, v_4 to be observed output variables), we obtain an L-CBN*

$$M = \left(G^+ = (J, (V, U), E^+), \left(P_v(X_v | X_{\mathrm{Pa}^{G^+}(v)})\right)_{v \in V \cup U}\right),$$

with $J = \{v_2\}$, $V = \{v_3, v_4\}$, $U = \{v_1\}$. Its induced observable CADMG $G := (G^+)^{\setminus \{v_1\}}$ is displayed in Figure 10(b). Its observable Markov kernel is the marginal $P(X_3, X_4 | \operatorname{do}(X_2))$ of the Markov kernel:

$$P(X_1, X_3, X_4 | \operatorname{do}(X_2)) = P_4(X_4 | X_1, X_2) \otimes P_2(X_3 | X_1, X_2) \otimes P_1(X_1).$$

4.2. Global Markov Property

Theorem 4.2.1 (Global Markov property for causal Bayesian networks)**.** *Consider a causal Bayesian network M with observable CADMG $G = (J, V, E, L)$ and observable Markov kernel $P(X_V | \operatorname{do}(X_J))$. Then for all $A, B, C \subseteq J \cup V$ (not necessarily disjoint) we have the implication:*

$$A \underset{G}{\overset{d}{\perp}} B \,|\, C \qquad \Longrightarrow \qquad X_A \underset{P(X_V | \operatorname{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C.$$

Remark 4.2.2. *If one wants to make the implicit dependence on J in Theorem 4.2.1 more explicit one can equivalently also write:*

$$A \underset{G}{\overset{d}{\perp}} J \cup B \,|\, C \qquad \Longrightarrow \qquad X_A \underset{P(X_V | \operatorname{do}(X_J))}{\perp\!\!\!\perp} X_J, X_B \,|\, X_C.$$

Notation 4.2.3. *Let $A, B, C \subseteq J \cup V$ with $X_A \perp\!\!\!\perp_{P(X_V | \operatorname{do}(X_J))} X_B \,|\, X_C$, then we have a factorization:*

$$P(X_A, X_B, X_C | \operatorname{do}(X_J)) = Q(X_A | X_C) \otimes P(X_B, X_C | \operatorname{do}(X_J)),$$

for some Markov kernel: $Q(X_A | X_C)$. If we marginalize out X_B and the deterministic $X_{C \cap J}$, we get:

$$P(X_A, X_{C \cap V} | \operatorname{do}(X_J)) = Q(X_A | X_C) \otimes P(X_{C \cap V} | \operatorname{do}(X_J)).$$

So we see that $Q(X_A|X_C)$ is a conditional Markov kernel:

$$P(X_A|X_{C\cap V}, \mathrm{do}(X_J))$$

that does only depend on $X_{J\cap C}$ in the do-*part. So we will use the following notation for $Q(X_A|X_C)$ (or in any other order behind the conditioning line):*

$$P(X_A|X_{C\cap V}, \mathrm{do}(X_{C\cap J}), \cancel{\mathrm{do}(X_J)}) := Q(X_A|X_C).$$

Note that by Theorem 2.5.28 we may (but do not need to) explicitly mention X_B as in:

$$P(X_A|\cancel{X_B}, X_{C\cap V}, \mathrm{do}(X_{C\cap J}), \cancel{\mathrm{do}(X_J)}),$$

because the Markov kernels are almost surely equal:

$$P(X_A|X_{C\cap V}, \mathrm{do}(X_{C\cap J}), \cancel{\mathrm{do}(X_J)}) = P(X_A|\cancel{X_B}, X_{C\cap V}, \mathrm{do}(X_{C\cap J}), \cancel{\mathrm{do}(X_J)}) \qquad P(X_C|X_J)\text{-a.s.}.$$

In these suggestive notations we can state the global Markov property (Theorem 4.2.1) as:

$$\begin{aligned} & A \overset{d}{\underset{G}{\perp}} B \,|\, C \\ & \implies P(X_A|X_B, X_C, \mathrm{do}(X_J)) \\ & \qquad = P(X_A|\cancel{X_B}, X_{C\cap V}, \mathrm{do}(X_{C\cap J}), \cancel{\mathrm{do}(X_J)})) \qquad P(X_B, X_C|\,\mathrm{do}(X_J))\text{-a.s.} \\ & \qquad = P(X_A|X_{C\cap V}, \mathrm{do}(X_{C\cap J}), \cancel{\mathrm{do}(X_J)}) \qquad P(X_B, X_C|\,\mathrm{do}(X_J))\text{-a.s.}. \end{aligned}$$

Proofs - Global Markov Property

The proof of the global Markov property follows similar arguments as used in [LDLL90, Ver93, Ric03, FM17, FM18, RERS23], namely chaining the separoid axioms together in an inductive way. The main difference here is that we never rely on the Symmetry property but instead use the left and right versions of the separoid axioms separately.

Theorem 4.2.4 (Global Markov property for causal Bayesian networks)**.** *Consider a causal Bayesian network M with observable CADMG $G = (J, V, E, L)$ and observable Markov kernel $P(X_V|\,\mathrm{do}(X_J))$. Then for all $A, B, C \subseteq J \cup V$ (not-necessarily disjoint) we have the implication:*

$$A \overset{d}{\underset{G}{\perp}} B \,|\, C \qquad \implies \qquad X_A \underset{P(X_V|\,\mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C.$$

If one wants to make the implicit dependence on J more explicit one can equivalently also write:

$$A \overset{d}{\underset{G}{\perp}} J \cup B \,|\, C \qquad \implies \qquad X_A \underset{P(X_V|\,\mathrm{do}(X_J))}{\perp\!\!\!\perp} X_J, X_B \,|\, X_C.$$

Proof. Because d-separation is preserved under marginalization:

$$A \underset{G}{\overset{d}{\perp}} B \,|\, C \quad\iff\quad A \underset{G^+}{\overset{d}{\perp}} B \,|\, C,$$

we can directly assume that we work with the causal Bayesian network without latent variables that marginalizes to the given one. So w.l.o.g. $L = \emptyset$ and G is a CDAG.
We then do induction by $\#V$.

0.) Induction start: $V = \emptyset$. This means that $A, B, C \subseteq J$. The assumption:

$$A \underset{G}{\overset{d}{\perp}} B \,|\, C,$$

implies that we must have that $A \subseteq C$. Otherwise a trivial walk from $A \subseteq J$ to $J \cup B$ would be C-open. Since $A, B, C \subseteq J$ we have the factorization:

$$P(X_A, X_B, X_C |\, \mathrm{do}(X_J)) = \underbrace{\bigotimes_{w \in A} \delta(X_w|X_w)}_{=:Q(X_A|X_C)} \otimes \underbrace{\bigotimes_{w \in B} \delta(X_w|X_w) \otimes \bigotimes_{w \in C} \delta(X_w|X_w)}_{=P(X_B, X_C |\, \mathrm{do}(X_J))}.$$

Because $A \subseteq C$ the Markov kernel $Q(X_A|X_C) := \bigotimes_{w \in A} \delta(X_w|X_w)$ really is a Markov kernel from $\mathcal{X}_C \dashrightarrow \mathcal{X}_A$. This already shows:

$$X_A \underset{P(X_V |\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B | X_C.$$

(IND): Induction assumption: The global Markov property holds for all causal Bayesian networks (with input variables, but without latent variables and without bi-directed edges) with $\#V < n$ (and arbitrary J).

1.) Now assume: $\#V = n > 0$ and $A \perp^d_G B \,|\, C$.
Since G is acyclic we can find a topological order $<$ for G where the elements of J are ordered first. Let $v \in V$ be its last element, which is thus childless.
Note that, since $\mathrm{Ch}^G(v) = \emptyset$, the marginalization $G^{\setminus\{v\}}$ has no bi-directed edges and thus induces again a causal Bayesian network without latent variables with $\#V^{\setminus\{v\}} = n - 1 < n$.
Furthermore, we have the factorization:

$$P(X_V |\, \mathrm{do}(X_J)) = P_v(X_v | X_{\mathrm{Pa}^G(v)}) \otimes \underbrace{\bigotimes_{w \in \mathrm{Pred}^G_<(v) \setminus J} P_w(X_w | X_{\mathrm{Pa}^G(w)})}_{P(X_{\mathrm{Pred}^G_<(v) \setminus J} |\, \mathrm{do}(X_J))}.$$

This factorization implies that we already have the conditional independence:

$$X_v \underset{P(X_V |\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{\mathrm{Pred}^G_<(v)} \,|\, X_D,$$

where we put $D := \mathrm{Pa}^G(v)$.

In the following we will distinguish between 4 cases:

A.) $v \in A \setminus C$,

B.) $v \in B \setminus C$,

C.) $v \in C$,

D.) $v \notin A \cup J \cup B \cup C$,

Note that $v \in V$, thus $v \notin J$, which shows that the above cover all possible cases. Further note that:

$$A \overset{d}{\underset{G}{\perp}} B \,|\, C,$$

implies that:

$$A \cap (J \cup B) \subseteq C.$$

Otherwise a trivial walk from A to $J \cup B$ would be C-open. This shows that $A \setminus C$, $(J \cup B) \setminus C$ and C are pairwise disjoint.

Case D.): $v \notin A \cup J \cup B \cup C$. Then we can marginalize out v and use the equivalence:

$$A \overset{d}{\underset{G}{\perp}} B \,|\, C \quad \iff \quad A \overset{d}{\underset{G^{\setminus v}}{\perp}} B \,|\, C.$$

With $\#V^{\setminus\{v\}} < n$ and induction (IND) we then get:

$$X_A \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C.$$

This shows the claim in case D.

Case A.): $v \in A \setminus C$. Then we can write:

$$\begin{aligned} A &= A' \,\dot\cup\, (A \cap C) \,\dot\cup\, \{v\}, \\ B &= B' \,\dot\cup\, (B \cap C), \end{aligned}$$

with some disjoint $A' \subseteq A \setminus C$ and $B' \subseteq B \setminus C$. We then have the implications:

$$\begin{aligned} A \overset{d}{\underset{G}{\perp}} B \,|\, C \xRightarrow{\text{Right Decomposition}} & \quad A \overset{d}{\underset{G}{\perp}} B' \,|\, C \\ \xRightarrow{\text{Left Decomposition}} & \quad A' \overset{d}{\underset{G}{\perp}} B' \,|\, C \\ \xRightarrow{\text{marginalization, } v \notin A' \cup J \cup B' \cup C} & \quad A' \overset{d}{\underset{G^{\setminus\{v\}}}{\perp}} B' \,|\, C \\ \xRightarrow{\text{induction (IND)}} & \quad X_{A'} \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{B'} \,|\, X_C. \qquad (\#1) \end{aligned}$$

On the other hand we have with $D = \mathrm{Pa}^G(v)$:

$$\begin{aligned}
A \overset{d}{\underset{G}{\perp}} B \,|\, C \quad & \xRightarrow{\text{Right Decomposition, } B' \subseteq B} & A \overset{d}{\underset{G}{\perp}} B' \,|\, C \\
& \xRightarrow{\text{Left Weak Union, } A = A' \,\dot\cup\, (A \cap C) \,\dot\cup\, \{v\}} & \{v\} \overset{d}{\underset{G}{\perp}} B' \,|\, A' \,\dot\cup\, C \\
& \xRightarrow{(*)\text{, see below}} & D \overset{d}{\underset{G}{\perp}} B' \,|\, A' \,\dot\cup\, C \\
& \xRightarrow{\text{marginalization, } v \notin D \cup J \cup B' \cup A' \cup C} & D \overset{d}{\underset{G \setminus \{v\}}{\perp}} B' \,|\, A' \,\dot\cup\, C \\
& \xRightarrow{\text{induction (IND)}} & X_D \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{B'} \,|\, X_{A' \,\dot\cup\, C} \\
& \xRightarrow{A' \,\dot\cup\, C} & X_D \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{B'} \,|\, X_{A'}, X_C. \qquad (\#2)
\end{aligned}$$

(*) holds since every $(A' \,\dot\cup\, C)$-open walk $w \;{*\!\!-\!\!*}\; \cdots$ from a $w \in D = \mathrm{Pa}^G(v)$ to $J \cup B'$ extends to an $(A' \,\dot\cup\, C)$-open walk from v to $J \cup B'$ via $v \leftarrow w \;{*\!\!-\!\!*}\; \cdots$, as w stays a non-collider in the extended walk (not in $A' \,\dot\cup\, C$) and $v \notin A' \,\dot\cup\, C$.

As discussed above we also already have the conditional independence:

$$X_v \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{\mathrm{Pred}^G_<(v)} \,|\, X_D.$$

With this and $A' \,\dot\cup\, B' \,\dot\cup\, C \subseteq \mathrm{Pred}^G_<(v)$ we get the implications:

$$\begin{aligned}
& & X_v \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{\mathrm{Pred}^G_<(v)} \,|\, X_D \\
& \xRightarrow{\text{Right Decomposition}} & X_v \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{A'}, X_{B'}, X_C \,|\, X_D \\
& \xRightarrow{\text{Right Weak Union}} & X_v \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{B'} \,|\, X_{A'}, X_C, X_D \\
& \xRightarrow{\text{Left Contraction, } (\#2)} & X_v, X_D \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{B'} \,|\, X_{A'}, X_C \\
& \xRightarrow{\text{Left Decomposition}} & X_v \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{B'} \,|\, X_{A'}, X_C \\
& \xRightarrow{\text{Left Contraction, } (\#1)} & X_{A'}, X_v \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{B'} \,|\, X_C \\
& \xRightarrow{X_J\text{-Inverted Right Decomposition}} & X_{A'}, X_v \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_J, X_{B'}, X_C \,|\, X_C \\
& \xRightarrow{\text{Right Decompositon, } B \subseteq B' \,\dot\cup\, C} & X_{A'}, X_v \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C. \qquad (\#3)
\end{aligned}$$

By (Extended) Left Redundancy we have:

$$X_{A'}, X_v, X_C \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_{A'}, X_v, X_C.$$

With this we get the implications:

$$
\begin{aligned}
& && X_{A'}, X_v, X_C \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \mid X_{A'}, X_v, X_C \\
&\xRightarrow{\text{Left Contraction, (\#3)}} && X_{A'}, X_v, X_{A'}, X_v, X_C \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \mid X_C \\
&\xRightarrow{\text{Left Decomposition, } A \subseteq A' \dot\cup \{v\} \dot\cup C} && X_A \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \mid X_C.
\end{aligned}
$$

This shows the claim in case A.

Case B.): $v \in B \setminus C$. Then we can write:

$$
\begin{aligned}
A &= A' \dot\cup (A \cap C), \\
B &= B' \dot\cup (B \cap C) \dot\cup \{v\},
\end{aligned}
$$

with some disjoint $A' \subseteq A \setminus C$ and $B' \subseteq B \setminus C$.

We then have the implications:

$$
\begin{aligned}
A \overset{d}{\underset{G}{\perp}} B \mid C &\xRightarrow{\text{Left Decomposition}} && A' \overset{d}{\underset{G}{\perp}} B \mid C \\
&\xRightarrow{\text{Right Decomposition}} && A' \overset{d}{\underset{G}{\perp}} B' \mid C \\
&\xRightarrow{\text{marginalization, } v \notin A' \cup J \cup B' \cup C} && A' \overset{d}{\underset{G \setminus \{v\}}{\perp}} B' \mid C \\
&\xRightarrow{\text{induction (IND)}} && X_{A'} \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{B'} \mid X_C. \qquad (\#1')
\end{aligned}
$$

Again with $D = \mathrm{Pa}^G(v)$ we get:

$$
\begin{aligned}
A \overset{d}{\underset{G}{\perp}} B \mid C &\xRightarrow{\text{Left Decomposition}} && A' \overset{d}{\underset{G}{\perp}} B \mid C \\
&\xRightarrow{\text{Right Decomposition}} && A' \overset{d}{\underset{G}{\perp}} B' \cup \{v\} \mid C \\
&\xRightarrow{\text{Right Weak Union}} && A' \overset{d}{\underset{G}{\perp}} \{v\} \mid B' \dot\cup C \\
&\xRightarrow{(\bullet)\text{, see below}} && A' \overset{d}{\underset{G}{\perp}} D \mid B' \dot\cup C \\
&\xRightarrow{\text{marginalization, } v \notin A' \cup J \cup D \cup B' \cup C} && A' \overset{d}{\underset{G \setminus \{v\}}{\perp}} D \mid B' \dot\cup C \\
&\xRightarrow{\text{induction (IND)}} && X_{A'} \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_D \mid X_{B' \dot\cup C} \\
&\xRightarrow{B' \dot\cup C} && X_{A'} \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_D \mid X_{B'}, X_C. \qquad (\#2')
\end{aligned}
$$

($\bullet$) holds since every $(B' \dot\cup C)$-open walk $\cdots \ast\!\!-\!\!\ast\, w$ from A' to a $w \in J \cup D$ extends to a $(B' \dot\cup C)$-open walk from A' to $J \cup \{v\}$, either because $w \in J$ or via $\cdots \ast\!\!-\!\!\ast\, w \longrightarrow v$ if $w \in D = \mathrm{Pa}^G(v)$. Note again that w stays a non-collider in the extended walk (outside of $B' \dot\cup C$) and $v \notin B' \dot\cup C$.

As before we will use the following conditional independence:

$$X_v \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{\mathrm{Pred}^G_<(v)} \,|\, X_D.$$

With this and $A' \cup J \cup B' \cup C \subseteq \mathrm{Pred}^G_<(v)$ we get the implications:

$$\begin{aligned}
& & X_v \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{\mathrm{Pred}^G_<(v)} \,|\, X_D \\
&\xRightarrow{\text{Right Decomposition}} & X_v \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{A'}, X_{B'}, X_C \,|\, X_D \\
&\xRightarrow{\text{Right Weak Union}} & X_v \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{A'} \,|\, X_{B'}, X_C, X_D \\
&\xRightarrow{\text{Flipped Left Cross Contraction, (\#2')}} & X_{A'} \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_D, X_v \,|\, X_{B'}, X_C \\
&\xRightarrow{\text{Right Decomposition}} & X_{A'} \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_v \,|\, X_{B'}, X_C \\
&\xRightarrow{\text{Right Contraction, (\#1')}} & X_{A'} \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{B'}, X_v \,|\, X_C \\
&\xRightarrow{X_J\text{-Inverted Right Decomposition}} & X_{A'} \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_J, X_{B'}, X_v, X_C \,|\, X_C \\
&\xRightarrow{\text{Right Decomposition, } B \subseteq B' \dot\cup \{v\} \dot\cup C} & X_{A'} \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C. \qquad (\#3')
\end{aligned}$$

By Redundancy we have:

$$X_{A'}, X_C \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_{A'}, X_C.$$

With this we get the implications:

$$\begin{aligned}
& & X_{A'}, X_C \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_{A'}, X_C \\
&\xRightarrow{\text{Left Contraction, (\#3')}} & X_{A'}, X_{A'}, X_C \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C. \\
&\xRightarrow{\text{Left Decomposition, } A \subseteq A' \dot\cup C} & X_A \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C.
\end{aligned}$$

This shows the claim in case B.

Case C.): $v \in C$. Then we can write:

$$\begin{aligned}
A &= A' \dot\cup (A \cap C), \\
B &= B' \dot\cup (B \cap C), \\
C &= C' \dot\cup \{v\},
\end{aligned}$$

with some pairwise disjoint $A' \subseteq A \setminus C$, $B' \subseteq B \setminus C$ and $C' \subseteq C$.

We then get the implications.

$$
\begin{aligned}
A \overset{d}{\underset{G}{\perp}} B \,|\, C &\xRightarrow{\text{Left Decomposition}} & A' \overset{d}{\underset{G}{\perp}} B \,|\, C \\
&\xRightarrow{\text{Right Decomposition}} & A' \overset{d}{\underset{G}{\perp}} B' \,|\, C \\
&\xRightarrow{C=C' \dot\cup \{v\}} & A' \overset{d}{\underset{G}{\perp}} B' \,|\, C' \dot\cup \{v\}
\end{aligned}
$$

We now claim that:

$$A' \overset{d}{\underset{G}{\perp}} B' \,|\, C' \dot\cup \{v\}$$

implies that one of the following statements holds:

$$A' \dot\cup \{v\} \overset{d}{\underset{G}{\perp}} B' \,|\, C' \qquad \vee \qquad A' \overset{d}{\underset{G}{\perp}} B' \dot\cup \{v\} \,|\, C'.$$

Assume the contrary:

$$A' \dot\cup \{v\} \overset{d}{\underset{G}{\not\perp}} B' \,|\, C' \qquad \wedge \qquad A' \overset{d}{\underset{G}{\not\perp}} B' \dot\cup \{v\} \,|\, C'.$$

So there exist shortest C'-open walks π_1 and π_2 in G such that all colliders are in C':

$$\pi_1 : \quad A' \cup \{v\} \ni u_0 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast u_k \in J \cup B',$$

and:

$$\pi_2 : \quad A' \ni w_0 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast w_m \in J \cup (B' \dot\cup \{v\}).$$

So all non-colliders of π_1 and π_2 are outside of C'. Since we consider shortest walks and $v \notin C'$ at most an end node of π_1 and π_2 could be equal to v. Otherwise one could shorten the walk.

Then note that $v \notin A'$ and $v \notin J \cup B'$, thus: $u_k \neq v$ and $w_0 \neq v$.

If now π_i does not contain v as an (end) node, then π_i would be $(C' \dot\cup \{v\})$-open, which is a contradiction to the assumption:

$$A' \overset{d}{\underset{G}{\perp}} B' \,|\, C' \dot\cup \{v\}.$$

So we can assume that the other end nodes equal v, i.e.: $u_0 = v$ and $w_m = v$.

Furthermore, both π_1 and π_2 are non-trivial walks, since $u_0 \neq u_k$ and $w_0 \neq w_m$. Since v is childless and $k, m \geq 1$ we have that the π_i are of the forms:

$$\pi_1 : \quad v \leftarrow u_1 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast u_k,$$

and:

$$\pi_2 : \quad w_0 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast w_{m-1} \rightarrow v,$$

with $u_1, w_{m-1} \in D = \mathrm{Pa}^G(v)$. Then the following walk:

$$A' \ni w_0 \ast\!\!\!-\!\!\!\ast \cdots \ast\!\!\!-\!\!\!\ast w_{m-1} \longrightarrow v \longleftarrow u_1 \ast\!\!\!-\!\!\!\ast \cdots \ast\!\!\!-\!\!\!\ast u_k \in J \cup B',$$

is a $(C' \dot\cup \{v\})$-open walk from A' to $J \cup B'$, in contradiction to:

$$A' \overset{d}{\underset{G}{\perp}} B' \,|\, C' \dot\cup \{v\}.$$

So the claim:

$$A' \dot\cup \{v\} \overset{d}{\underset{G}{\perp}} B' \,|\, C' \qquad \vee \qquad A' \overset{d}{\underset{G}{\perp}} B' \dot\cup \{v\} \,|\, C',$$

must be true. So we reduced case C to case A or case B, which then imply:

$$X_A, X_v \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_{C'} \qquad \vee \qquad X_A \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B, X_v \,|\, X_{C'}.$$

If we apply Left Weak Union to the left and Right Weak Union to the right we get:

$$X_A \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_{C'}, X_v,$$

which implies:

$$X_A \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C.$$

This shows the claim in case C. □

4.3. Operations on Causal Bayesian Networks

4.3.1. Hard Interventions on Causal Bayesian Networks

Definition 4.3.1 (Hard intervention on causal Bayesian network)**.** *Consider a causal Bayesian network (CBN) given by:*

$$M = \Big(G = (J, V, E), \big(P_v(X_v|X_{\mathrm{Pa}^G(v)})\big)_{v \in V}\Big).$$

Now let $W \subseteq J \cup V$ be any subset. Then we define the ***intervened causal Bayesian network*** *w.r.t. W via:*

1. *CDAG: $G_{\mathrm{do}(W)} = (J \cup W, V \setminus W, E_{\mathrm{do}(W)})$, and:*
2. *Markov kernels: $P_v(X_v|X_{\mathrm{Pa}^G(v)})$ for $v \in V \setminus W$.*

Its observable Markov kernel is then:

$$P(X_{V \setminus W} \,|\, \mathrm{do}(X_{J \cup W})) = \bigotimes_{v \in V \setminus W} P_v(X_v|X_{\mathrm{Pa}^G(v)}).$$

Note that if $v \in V \setminus W$ then $\mathrm{Pa}^G(v) = \mathrm{Pa}^{G_{\mathrm{do}(W)}}(v)$.

Remark 4.3.2. *Note that the above notations imply for every $v \in V$ and $W \subseteq V \setminus \{v\}$ the identifications:*

$$P_v(X_v|X_{\mathrm{Pa}^G(v)}) = P(X_v|\operatorname{do}(X_{J\cup V\setminus W})) = P(X_v|\operatorname{do}(X_{\mathrm{Pa}^G(v)})),$$

which we will use interchangably in the following.

Remark 4.3.3 (Hard intervention on causal Bayesian network with latent variables)**.** *We define hard interventions on an L-CBN the same way as on a CBN, but we usually only allow for interventions on sets $W \subseteq J \cup V$, i.e. with $W \cap U = \emptyset$, where U is the set of latent variables.*

4.3.2. Node-Splitting on Causal Bayesian Networks

Definition 4.3.4 (Node-splitting on causal Bayesian network)**.** *Consider a causal Bayesian network (CBN) given by $(G, P(X_V|\operatorname{do}(X_J)))$ with CDAG: $G = (J, V, E)$ and Markov kernels: $P_v(X_v|X_{\mathrm{Pa}^G(v)})$ for $v \in V$. Now let $W \subseteq V$ be any subset. Then we define the* ***node-split CBN w.r.t.*** *W as the causal Bayesian network given by:*

1. *CDAG: $G' := G_{\mathrm{split}(W)} = (J, (V \setminus W) \cup W^0 \cup W^1, E_{\mathrm{split}(W)})$, and:*
2. *Markov kernels for $v \in V$:*

$$P_{v^0}(X_{v^o} \in A|X_{\mathrm{Pa}^{G'}(v^o)} = \tilde{x}) := P_v(X_v \in A|X_{\mathrm{Pa}^G(v)} = \tilde{x}),$$

where for brevity we put $v^o := v$ for $v \in V \setminus W$, and for $w^1 \in W^1$:

$$P_{w^1}(X_{w^1} \in A|X_{w^0} = \tilde{x}) := \delta(X_{w^1} \in A|X_{w^0} = \tilde{x}).$$

Remark 4.3.5. *Similarly, we can define node-splitting interventions on causal Bayesian network with latent variables, but allow only W with $W \cap U = \emptyset$.*

This operation can be used to reason about "single-world" counterfactuals using CBNs. We will not dwell on the details here, but discuss this in Chapter 8 in the context of SCMs.

4.3.3. Node-Splitting Hard Interventions on Causal Bayesian Networks

Definition 4.3.6 (Node-splitting hard intervention on causal Bayesian network)**.** *Consider a causal Bayesian network (CBN) given by $(G, P(X_V|\operatorname{do}(X_J)))$ with CDAG: $G = (J, V, E)$ and Markov kernels: $P_v(X_v|X_{\mathrm{Pa}^G(v)})$ for $v \in V$. Now let $W \subseteq V$ be any subset. Then we define the* ***node-splitting hard intervention*** *w.r.t. W as the causal Bayesian network given by:*

1. *CDAG: $G' := G_{\mathrm{swig}(W)} = (J \,\dot{\cup}\, W^i, W^o \,\dot{\cup}\, V \setminus W, E_{\mathrm{swig}(W)})$, and:*
2. *Markov kernels for $v \in V$:*

$$P_{v^o}(X_{v^o} \in A|X_{\mathrm{Pa}^{G'}(v^o)} = \tilde{x}) := P_v(X_v \in A|X_{\mathrm{Pa}^G(v)} = \tilde{x}),$$

where for brevity we put $v^o := v$ for $v \in V \setminus W$.

Remark 4.3.7. *Similarly, we can define node-splitting hard interventions on causal Bayesian network with latent variables, but allow only W with $W \cap U = \emptyset$.*

4.3.4. Soft Interventions on Causal Bayesian Networks

Remark 4.3.8 (Modelling soft interventions on causal Bayesian networks)**.** *Consider a causal Bayesian network given by* $(G, P(X_V|\operatorname{do}(X_J)))$ *with CDAG:* $G = (J, V, E)$ *and Markov kernels:* $P_v(X_v|X_{\operatorname{Pa}^G(v)})$ *for* $v \in V$.
Let $W \subseteq J \cup V$. *In order to model a soft intervention on variables* X_w *for* $w \in W \setminus J$, *we introduce intervention nodes* $I_w \longrightarrow w$ *for* $w \in W \setminus J$, *which come with new input variables* X_{I_w}, *and replace the Markov kernel:*

$$P_w(X_w|X_{\operatorname{Pa}^G(w)})$$

for $w \in W \setminus J$ *by one that models the dependence on the soft intervention variables properly:*

$$P_w(X_w|X_{\operatorname{Pa}^G(w)}, X_{I_w}).$$

For $w \in J$, *we simply identify* I_w *with* w. *So the* ***softly intervened causal Bayesian network*** *w.r.t.* W *then has:*

1. *CDAG:* $G_{\operatorname{do}(I_W)} = (J \dot\cup \{I_w | w \in W \setminus J\}, V, E \dot\cup \{I_w \longrightarrow w | w \in W \setminus J\})$, *and:*
2. *Markov kernels:*

   $P_v(X_v|X_{\operatorname{Pa}^G(v)})$ *for* $v \in V \setminus W$, *and:*

   $P_w(X_w|X_{\operatorname{Pa}^G(w)}, X_{I_w})$ *for* $w \in W \setminus J$.

Note that $\operatorname{Pa}^{G_{\operatorname{do}(I_W)}}(w) = \operatorname{Pa}^G(w) \dot\cup \{I_w\}$ *for* $w \in W \setminus J$ *and* $\operatorname{Pa}^{G_{\operatorname{do}(I_W)}}(v) = \operatorname{Pa}^G(v)$ *for* $v \in V \setminus W$.

Remark 4.3.9 (Modelling hard interventions with intervention nodes)**.** *It is sometimes beneficial to model hard interventions with intervention nodes. Let the setting be like in Remark 4.3.8. When we model hard interventions with intervention nodes we make the further more specific choices for* $w \in W \setminus J$:

1. $\mathcal{X}_{I_w} := \mathcal{X}_w \dot\cup \{\star\}$,
2. $P_w(X_w \in A|X_{\operatorname{Pa}^G(w)} = x_{\operatorname{Pa}^G(w)}, X_{I_w} = x_{I_w}) :=$

$$\begin{cases} P_w(X_w \in A|X_{\operatorname{Pa}^G(w)} = x_{\operatorname{Pa}^G(w)}), & \text{if} \quad x_{I_w} = \star, \\ \delta(X_w \in A|X_w = x_{I_w}) = \mathbb{1}_A(x_{I_w}), & \text{if} \quad x_{I_w} \neq \star. \end{cases}$$

Note that the CDAG will then rather be: $G_{\operatorname{do}(I_W)}$ *in contrast to:* $G_{\operatorname{do}(W)}$.
The above choices reflect that if we put $X_{I_w} = \star$ *then no intervention occurs and the value of* X_w *is (probabilistically) determined using the usual Markov kernel. But if we put* $X_{I_w} = x_{I_w} \neq \star$ *then we change the value of* X_w *to* x_{I_w} *(with 100% probability) independent of the values of its parents. This is then similar to the hard intervention:* $\operatorname{do}(X_w = x_{I_w})$. *This allows us to model simultaneously the unintervened and an intervened version of the CBN with a single CBN.*

Remark 4.3.10. *Again, we can do all the above also with causal Bayesian network with latent variables, but allow only* W *with* $W \cap U = \emptyset$.

4.3.5. Marginalization of Causal Bayesian Networks

Definition 4.3.11 (Marginalization of causal Bayesian network with latent variables)**.** *Consider a causal Bayesian network with latent variables (L-CBN):*

$$M = \left(G^+ = (J, (V, U), E^+), \left(P_v(X_v|X_{\mathrm{Pa}^{G^+}(v)})\right)_{v \in V \cup U}\right),$$

Let $W \subseteq V$ be a subset. We then define the ***marginalized L-CBN*** *by just replacing V with $V \setminus W$ and U with $U \dot{\cup} W$. The Markov kernels P_v for $v \in V \dot{\cup} U = (V \setminus W) \dot{\cup} (U \dot{\cup} W)$ stay the same.*
With this definition the observable Markov kernel marginalizes to:

$$P\left(X_{V \setminus W} \middle| \operatorname{do}(X_J)\right),$$

and the observable CADMG becomes:

$$(G^+)^{\setminus (U \dot{\cup} W)} = G^{\setminus W},$$

i.e. the marginalized G w.r.t. W.

4.4. Standard Forms of Causal Bayesian Networks

Definition 4.4.1. *Consider two causal Bayesian network with latent variables (L-CBNs):*

$$M_1 = \left(G_1^+ = \left(J_1, (V_1, U_1), E_1^+\right), \left(P_{1,v}(X_v|X_{\mathrm{Pa}^{G_1^+}(v)})\right)_{v \in U_1 \cup V_1}\right),$$
$$M_2 = \left(G_2^+ = \left(J_2, (V_2, U_2), E_2^+\right), \left(P_{2,v}(X_v|X_{\mathrm{Pa}^{G_2^+}(v)})\right)_{v \in U_2 \cup V_2}\right).$$

We call them ***interventionally equivalent*** *if all of the following conditions hold:*

1. $J_1 = J_2 =: J$,
2. $V_1 = V_2 =: V$,
3. $\mathcal{X}_{1,v} = \mathcal{X}_{2,v} =: \mathcal{X}_v$ *for all* $v \in J \cup V$,
4. *for all subsets $W \subseteq V$ we have the equality of the intervened Markov kernels:*

$$P_1\left(X_{V \setminus W} \middle| \operatorname{do}(X_{J \cup W})\right) = P_2\left(X_{V \setminus W} \middle| \operatorname{do}(X_{J \cup W})\right).$$

Definition 4.4.2 (Cliques and maximal cliques of undirected graphs)**.** *Let $G = (V, L)$ be an undirected graph. A set of nodes $W \subseteq V$ is called a* ***clique***[17] *of G if for all $w_1, w_2 \in W$ with $w_1 \neq w_2$ we have that the edge $w_1 — w_2 \in L$. A clique W is called a* ***maximal clique*** *of G if for every clique $\tilde{W}$ of G with $W \subseteq \tilde{W}$ we have that $W = \tilde{W}$.*

[17] A clique is also called *complete subgraph* in the literature.

Definition/Theorem 4.4.3 (Standard forms of L-CBNs)**.** *Consider a causal Bayesian network M with latent variables (L-CBN) with observable CADMG $G = (J, V, E, L)$. Let*

$$\mathcal{C} := \{W \subseteq V \,|\, W \text{ maximal clique of } (V, L)\} .$$

the sets of all maximal cliques of the (undirected) graph consisting only of the nodes from V and the bi-directed edges from G. Define the set of (latent) nodes:

$$\tilde{U} := \{\tilde{u}_W \,|\, W \in \mathcal{C}\} ,$$

and directed edges:

$$\tilde{E}^+ := E \,\dot{\cup}\, \{\tilde{u}_W \longrightarrow w \,|\, W \in \mathcal{C}, w \in W\} .$$

Then there exists an L-CBN of the form:

$$\tilde{M} = \left(\tilde{G}^+ = \left(J, (V, \tilde{U}), \tilde{E}^+\right), \left(\tilde{P}_v(X_v|X_{\mathrm{Pa}^{\tilde{G}^+}(v)})\right)_{v \in V \cup \tilde{U}}\right)$$

*that is interventionally equivalent to M. Furthermore, we can choose to arrange them in **one** of the following ways:*

1. ***Structural causal model form**: All Markov kernels for $v \in V$ are deterministic:*

   $$\tilde{P}_v(X_v \in A|X_{\mathrm{Pa}^{G^+}(v)} = \tilde{x}) = \delta(R_v \in A|X_{\mathrm{Pa}^{G^+}(v)} = \tilde{x}),$$

   for some measurable maps R_v, $v \in V$. OR:

2. ***Canonical form**: All latent variables $\tilde{u}_W$ with $\#W = 1$ and the corresponding variables, edges and Markov kernels can be removed from $\tilde{M}$ as well, leaving us only with the latent variables $\tilde{u}_W$ with $W \in \mathcal{C}$ and $\#W \geq 2$.*

Remark 4.4.4. *Consider the standard forms $\tilde{M}$ of M from Definition/Theorem 4.4.3.*

1. *In particular, we have:*
   a) $(\tilde{G}^+)^{\setminus \tilde{U}} = G$,
   b) $\mathrm{Pa}^{\tilde{G}^+}(\tilde{U}) = \emptyset$,
   c) $\mathrm{Ch}^{\tilde{G}^+}(u) \in \mathcal{C}$ *for every* $u \in \tilde{U}$.

2. *We can use measurable embeddings/isomorphisms: $\mathcal{X}_u \hookrightarrow [0, 1]$ for $u \in \tilde{U}$ to further restrict to the case:*
   a) $\mathcal{X}_u \cong [0, 1]$,
   b) $\tilde{P}_u(X_u)$ *is the uniform distribution on* $[0, 1]$.

3. *Note that the Markov kernels dependent on $X_{\tilde{U}}$ might not be unique as we can always transform $[0, 1]$ to $[0, 1]$ in strange ways.*

4. *The construction of the canonical form generally*[18] *leads to an interventionally equivalent L-CBN with the smallest number of latent variables such that its observable CADMG stays unchanged.*

5. *The construction of the structural causal model form generally*[18] *leads to an interventionally equivalent L-CBN with the smallest number of latent variables such that its observable CADMG stays unchanged and such that every Markov kernel with non-trivial input is deterministic.*

Remark 4.4.5 (Marginalizations and hard interventions on standard forms)**.** *Let the following L-CBN be in one of the standard forms:*

$$\left(G^+ = (J, V, U, E^+),\ P\left(X_{V\cup U}\middle|\, \mathrm{do}\left(X_J\right)\right)\right).$$

Now let $W \subseteq V$ then we defined the marginalization w.r.t. W by replacing V with $V \setminus W$ and U with $U \cup W$. We could re-define the marginalization as a corresponding standard form of that procedure.
Similarly we could post-process hard interventions with standardization steps.

Proofs - Standard Forms of Causal Bayesian Networks

Proof. Step 1. For every $v \in V \cup U$ we can write the Markov kernel P_v as the composition of a deterministic one and a uniform distribution $P_{\bar{v}}(X_{\bar{v}})$ on $\mathcal{X}_{\bar{v}} := [0,1]$ by Remark 2.7.4:

$$P_v(X_v|X_{\mathrm{Pa}^{G^+}(v)}) = \delta(R_v|X_{\bar{v}}, X_{\mathrm{Pa}^{G^+}(v)}) \circ P_{\bar{v}}(X_{\bar{v}}).$$

We now put:

$$\bar{U} := U \,\dot{\cup}\, \{\bar{v}\,|\, v \in V \cup U\}, \qquad \bar{E}^+ := E^+ \,\dot{\cup}\, \{\bar{v} \longrightarrow v\,|\, v \in V \cup U\},$$

and to get $\bar{M}$ we add the $P_{\bar{v}}$ to M and replace P_v for $v \in V \cup U$ by the deterministic one given by:

$$\bar{P}_v(X_v \in A|X_{\bar{v}}, X_{\mathrm{Pa}^{G^+}(v)}) := \delta(R_v \in A|X_{\bar{v}}, X_{\mathrm{Pa}^{G^+}(v)}).$$

Then $\bar{G}^+$ clearly marginalizes to G^+ (when we marginalize out all the $\bar{v}$ again) and the marginal of:

$$\bar{P}_v(X_v \in A|X_{\bar{v}}, X_{\mathrm{Pa}^{G^+}(v)}) \otimes \bar{P}_{\bar{v}}(X_{\bar{v}}),$$

in the defining product of the joint Markov kernel is $P_v(X_v|X_{\mathrm{Pa}^{G^+}(v)})$ for all $v \in V \cup U$ by construction again.
Step 2. Marginalize out all $u \in U$. Let us first look at the Markov kernel side if we marginalize out X_u in the defining product of the joint Markov kernel for $u \in U$:

$$\begin{aligned}
&\int_{\mathcal{X}_u} \bigotimes_{v \in \mathrm{Ch}^{\bar{G}^+}(u)} \bar{P}_v(X_v|X_{\mathrm{Pa}^{\bar{G}^+}(v)\setminus\{u\}}, X_u = x_u)\, \delta(R_u \in dx_u|X_{\mathrm{Pa}^{\bar{G}^+}(u)}) \\
&= \bigotimes_{v \in \mathrm{Ch}^{\bar{G}^+}(u)} \bar{P}_v(X_v|X_{\mathrm{Pa}^{\bar{G}^+}(v)\setminus\{u\}}, X_u = R_u(X_{\mathrm{Pa}^{\bar{G}^+}(u)})),
\end{aligned}$$

[18] Excluding degenerate L-CBNs. In those cases one could possibly remove even more latent variables.

which is again a product (only) because we marginalized a deterministic Markov kernel out. So we define:

$$\hat{P}_v(X_v|X_{\mathrm{Pa}^{\hat{G}^+}(v)}) := \bar{P}_v(X_v|X_{\mathrm{Pa}^{\bar{G}^+}(v)\setminus\{u\}}, X_u = R_u(X_{\mathrm{Pa}^{\bar{G}^+}(u)})),$$

which is as the composition of deterministic Markov kernels again a deterministic Markov kernel. From this we also read off that we need to consider the graph $\hat{G}^+$ with:

$$\mathrm{Pa}^{\hat{G}^+}(v) := \mathrm{Pa}^{\bar{G}^+}(v) \setminus \{u\} \cup \mathrm{Pa}^{\bar{G}^+}(u),$$

i.e. the CDAG from $(\bar{G}^+)^{\setminus U}$ where we removed all bi-directed edges, and with latent nodes $\hat{U} = \bar{U} \setminus U$. Then note that for $u \in U$ we have:

$$\mathrm{Ch}^{\bar{G}^+}(u) = \mathrm{Ch}^{(\bar{G}^+)^{\setminus U}}(\bar{u}).$$

This implies that we recover the removed bi-directed edges if we further marginalize out all the $\bar{u}$, i.e.:

$$(\hat{G}^+)^{\setminus \hat{U}} = (\bar{G}^+)^{\setminus \bar{U}} = (G^+)^{\setminus U} = G.$$

Step 3. We marginalize out all nodes $u \in \hat{U}$ with $\mathrm{Ch}^{\hat{G}^+}(u) = \emptyset$. For those u we have:

$$\hat{P}(X_V, X_{\hat{U}}|\,\mathrm{do}(X_J)) = \hat{P}_u(X_u|X_{\mathrm{Pa}^{\hat{G}^+}(u)}) \otimes \hat{P}(X_V, X_{\hat{U}\setminus\{u\}}|\,\mathrm{do}(X_J)).$$

So marginalizing out X_u does not interfere with the rest of the Markov kernels. So from now on we can w.l.o.g. assume that $\#\mathrm{Ch}^{\hat{G}^+}(u) \geq 1$ for all $u \in \hat{U}$.
Step 4. Since $(\hat{G}^+)^{\setminus \hat{U}} = G$ we have that for each $u \in \hat{U}$ the set $\mathrm{Ch}^{\hat{G}^+}(u)$ is a clique of (V, L). So we can (arbitrarily) assign u to any maximal clique W of (V, L) with $\mathrm{Ch}^{\hat{G}^+}(u) \subseteq W$. So let W be a fixed maximal clique of (V, L) and $u_1, \dots, u_k \in \hat{U}$ be all $u \in \hat{U}$ that we assigned to W. Then we consider the space:

$$\mathcal{X}_{\tilde{u}_W} := \prod_{\ell=1}^{k} \mathcal{X}_{u_\ell},$$

and the variables:

$$X_{\tilde{u}_W} := (X_{u_\ell})_{\ell=1,\dots,k}.$$

Then every Markov kernel dependent on such an X_{u_ℓ} can be written as a Markov kernel dependent on $X_{\tilde{u}_W}$, by only using the u_ℓ component. We will then replace $u_1, \dots, u_k$ by the single node $\tilde{u}_W$ and every edge of form $u_\ell \longrightarrow v$ by $\tilde{u}_W \longrightarrow v$. If we do this for all $u \in \hat{U}$ and maximal cliques W of (V, L) we arrive at the CADMG $\tilde{G}^+ = (J, V, \tilde{U}, \tilde{E}^+)$, with:

$$\tilde{E}^+ := E \,\dot{\cup}\, \{\tilde{u}_W \longrightarrow w \,|\, W \in \mathcal{C}, w \in W\}.$$

So we arrived at the desired structural causal model form and one can convince oneself that at each step we get an interventionally equivalent L-CBN to the step before.

The canonical form follows from the structural causal model form by marginalizing out all X_u with $\#\mathrm{Ch}^{\tilde{G}^+}(u) \leq 1$, i.e. by replacing the left (deterministic) Markov kernel dependent on X_u in the product:

$$P_v(X_v | X_{\mathrm{Pa}^{\tilde{G}^+}(v) \setminus \{u\}}, X_u) \otimes P_u(X_u),$$

by the composition:

$$P_v(X_v | X_{\mathrm{Pa}^{\tilde{G}^+}(v) \setminus \{u\}}, X_u) \circ P_u(X_u),$$

which then might not be deterministic anymore. $\square$

5. Identification of Causal Effects in CBNs

This section investigates under which circumstances one can *identify causal effects* and estimate them just from observational data alone under the (strong) assumption that the underlying causal graph is known. More generally, we ask the question when an interventional Markov kernel of a causal Bayesian network can be identified from the causal graph G and the observational Markov kernel alone.

We will see that the main tool to allow for such statements is the global Markov property, see Theorem 4.2.1, applied to the causal Bayesian network that is augmented with further intervention variables.

We first study under which graphical conditions interventions don't have an effect or when one essentially can replace interventions with conditioning operations. These rules will be summarized as the *three rules of do-calculus*. The main references are [Pea93a, Pea93b, Pea09], also see [Pea95, FM20, For21].

We then study under which graphical criteria one gets explicit *adjustment formulas* to estimate interventional Markov kernels from observational ones. The literature mentions the *backdoor criterion*, see [Pea93a, Pea93b, Pea09], the *extended backdoor criterion*, see [PP10, SdWR10], the *selection backdoor criterion*, see [BTP14], criteria for *selection without/partial external data*, see [CB17, CTB18], and all their generalizations to the cyclic case, see [FM20], also see [SP06a, PTKM15, For21].

Finally, we present the *ID-algorithm*, which can decide just by processing the causal graph G if an interventional Markov kernel can be identified by the observational one (under further assumptions, like strict positivity, etc.). If the algorithm does not output FAIL then it also presents a formula to estimate the queried interventional Markov kernel. The main references for the ID-algorithm are [Pea09, GP95, Tia02, TP02, Tia04, SP06b, HV06, HV08, RERS23, FM20].

5.1. Do-Calculus

Remark 5.1.1 (Recap)**.** *Consider an L-CBN:*

$$M = \left(G^+ = \left(J, (V, U), E^+\right), \left(P_v(X_v|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)}))\right)_{v\in V\cup U}\right).$$

Then we get the joint Markov kernel over all input, observed and unobserved output variables as follows:

$$P\left(X_V, X_U, X_J|\,\mathrm{do}(X_J)\right) := \bigotimes_{v\in U\cup V} P_v\left(X_v|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)})\right) \otimes \bigotimes_{j\in J} \delta(X_j|X_j).$$

Further, for $D \subseteq J \cup V$ and $B \subseteq V \setminus D$ we get the combined hard and soft interventions:

$$P\left(X_{V\setminus D}, X_U, X_{J\cup D}, X_{I_B}|\,\mathrm{do}(X_{J\cup I_B\cup D})\right) :=$$

$$\bigotimes_{v\in V\setminus(B\cup D)} P_v\left(X_v|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)})\right) \otimes \bigotimes_{v\in B} P_v\left(X_v|\,\mathrm{do}(X_{\{I_v\}\cup\mathrm{Pa}^{G^+}(v)})\right) \otimes$$

$$\bigotimes_{v\in U} P_v\left(X_v|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)})\right)\otimes\bigotimes_{j\in J\cup D}\delta(X_j|X_j)\otimes\bigotimes_{v\in B}\delta(X_{I_v}|X_{I_v}),$$

where we need to reorder all the factors such that the product is in reverse order of a topological order and where we use the following Markov kernels to model hard interventions as soft interventions, $v\in B$:

$$P_v\left(X_v|\,\mathrm{do}\left(X_{\mathrm{Pa}^{G^+}(v)},X_{I_v}=x_{I_v}\right)\right):=\begin{cases}P_v\left(X_v|\,\mathrm{do}\left(X_{\mathrm{Pa}^{G^+}(v)}\right)\right), & \text{if}\quad x_{I_v}=\star,\\ \delta(X_v|X_v=x_{I_v}), & \text{if}\quad x_{I_v}\neq\star.\end{cases}$$

Finally we can also marginalize (i.e. integrating out) and condition to get:

$$P\left(X_{A\cup B}|X_C,\mathrm{do}(X_{J\cup I_B\cup D})\right),$$

for any $A,B,C\subseteq V$, $D\subseteq V\cup J$ with A,B,C,D disjoint (w.l.o.g. we can assume $A\subseteq V$, $B\cap C=\emptyset$ and $B\cap D=\emptyset$). For more suggestive formulas later on we also freely permute the order of symbols behind the conditioning line, e.g. (if $F\cap D=\emptyset$):

$$P\left(X_A|\,\mathrm{do}(X_F),X_C,\mathrm{do}(X_D)\right):=P\left(X_A|X_C,\mathrm{do}(X_D,X_F)\right):=P\left(X_A|X_C,\mathrm{do}(X_{D\cup F})\right).$$

Please note that no matter in which order we write the do*-part and conditioning part behind the conditioning line $|$, we always assume that we perform the intervention (*do*) first and afterwards condition.*
We will also make use of the following CADMG (for $I_B\cap D=\emptyset$):

$$G_{\mathrm{do}(I_B,D)}:=G_{\mathrm{do}(I_B\cup D)}=(G^+_{\mathrm{do}(I_B\cup D)})^{\setminus U}.$$

Theorem 5.1.2. *[Almost-sure do-calculus—in detail] Consider an L-CBN:*

$$M=\left(G^+=\left(J,(V,U),E^+\right),\left(P_v(X_v|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)}))\right)_{v\in V\cup U}\right).$$

Let $A,B,C\subseteq V$ and $D\subseteq J\cup V$ be such that A,B,C,D are pairwise disjoint.
Further assume that we have reference measures μ_v on $\mathcal{X}_v$ for every $v\in V$ that are each equivalent to a probability measure (in terms of absolute continuity).[19] *We then put $\mu_F:=\bigotimes_{v\in F}\mu_v$ for $F\subseteq V$.*

1. *Insertion/deletion of observation: Assume:*

$$A\underset{G_{\mathrm{do}(D)}}{\overset{d}{\perp}}B\,|\,C\cup D.$$

For a fixed finite index set I consider subsets $B^{(i)}\subseteq B$, for $i\in I$, and pick for each $i\in I$ an arbitrary version of a conditional Markov kernel:

$$P\left(X_A|X_{B^{(i)}},X_C,\mathrm{do}(X_{D\cup J})\right):\ \mathcal{X}_{B\cup C\cup D\cup J}\to\mathcal{X}_{B^{(i)}\cup C\cup D\cup J}\to\mathcal{P}(\mathcal{X}_A),$$

of $P(X_A,X_{B^{(i)}},X_C|\,\mathrm{do}(X_{D\cup J}))$. Then there exists a measurable $P(X_B,X_C|\,\mathrm{do}(X_{D\cup J}))$-null set $N\subseteq\mathcal{X}_{B\cup C\cup D\cup J}$, such that all those Markov kernels are equal on the complement N^{c}.

[19] Recall the connection between absolute continuity and strictly positive densities in Corollary 2.3.20. All σ-finite measures satisfy this assumption.

Note that if $\mu_{B\cup C} \ll P(X_B, X_C|\,\mathrm{do}(X_{D\cup J}))$ then N is also a $\mu_{B\cup C}$-null set, i.e. for every $x_{D\cup J} \in \mathcal{X}_{D\cup J}$ we have: $\mu_{B\cup C}(N_{x_{D\cup J}}) = 0$.

If we also have the reverse $P(X_B, X_C|\,\mathrm{do}(X_{D\cup J})) \ll \mu_{B\cup C}$ then we can change the above conditional Markov kernels on a $\mu_{B\cup C}$-null set N while they remain versions of the corresponding conditional Markov kernel.[20]

2. *Action/observation exchange: Assume:*

$$A \underset{G_{\mathrm{do}(I_B\cup D)}}{\overset{d}{\perp}} I_B \,|\, B\cup C\cup D.$$

For a fixed finite index set I consider decompositions $B = B_1^{(i)} \dot\cup B_2^{(i)}$, for $i \in I$, and pick for each $i \in I$ an arbitrary version of a conditional Markov kernel:

$$P(X_A|X_{B_1^{(i)}}, \mathrm{do}(X_{B_2^{(i)}}), X_C, \mathrm{do}(X_{D\cup J})):\ \mathcal{X}_{B\cup C\cup D\cup J} \to \mathcal{P}(\mathcal{X}_A),$$

of $P(X_A, X_{B_1^{(i)}}, X_C|\,\mathrm{do}(X_{B_2^{(i)}}, X_{D\cup J}))\otimes\mu_{B_2^{(i)}}$ and assume the following absolute continuities:

$$\mu_{B\cup C} \ll P(X_{B_1^{(i)}}, X_C|\,\mathrm{do}(X_{B_2^{(i)}}, X_{D\cup J})) \otimes \mu_{B_2^{(i)}}$$

for all $i \in I$.[21] *Then there exists a measurable $\mu_{B\cup C}$-null set $N \subseteq \mathcal{X}_{B\cup C\cup D\cup J}$, such that all those conditional Markov kernels are equal on the complement N^{c}.*

If we also assume the reverse absolute continuities for all $i \in I$:

$$P(X_{B_1^{(i)}}, X_C|\,\mathrm{do}(X_{B_2^{(i)}}, X_{D\cup J})) \otimes \mu_{B_2^{(i)}} \ll \mu_{B\cup C},$$

then all those conditional Markov kernels are versions of each other.[22]

[20] Note that the absolute continuities: $\mu_{B\cup C} \ll P(X_B, X_C|\,\mathrm{do}(X_{D\cup J})) \ll \mu_{B\cup C}$ hold if $P(X_B, X_C|\,\mathrm{do}(X_{D\cup J}))$ has a strictly positive Doob-Radon-Nikodym derivative w.r.t. $\mu_{B\cup C}$. Furthermore, the converse is also true for σ-finite reference measures $\mu_{B\cup C}$ by Corollary 2.3.20.

[21] If you instead expected to pick for each $i \in I$ an arbitrary version of a conditional Markov kernel:

$$P(X_A|X_{B_1^{(i)}}, \mathrm{do}(X_{B_2^{(i)}}), X_C, \mathrm{do}(X_{D\cup J})):\ \mathcal{X}_{B\cup C\cup D\cup J} \to \mathcal{P}(\mathcal{X}_A),$$

of $P(X_A, X_{B_1^{(i)}}, X_C|\,\mathrm{do}(X_{B_2^{(i)}}, X_{D\cup J}))$ and to assume the absolute continuities

$$\mu_{B_1^{(i)}\cup C} \ll P(X_{B_1^{(i)}}, X_C|\,\mathrm{do}(X_{B_2^{(i)}}, X_{D\cup J})) \ll \mu_{B_1^{(i)}\cup C}$$

for all $i \in I$: that would lead to a similar, but slightly weaker statement.

[22] Note that the absolute continuities: $\mu_{B\cup C} \ll P(X_{B_1^{(i)}}, X_C|\,\mathrm{do}(X_{B_2^{(i)}}, X_{D\cup J})) \otimes \mu_{B_2^{(i)}} \ll \mu_{B\cup C}$ hold if the absolute continuities: $\mu_{B_1^{(i)}\cup C} \ll P(X_{B_1^{(i)}}, X_C|\,\mathrm{do}(X_{B_2^{(i)}}, X_{D\cup J})) \ll \mu_{B_1^{(i)}\cup C}$ hold, which hold if $P(X_{B_1^{(i)}}, X_C|\,\mathrm{do}(X_{B_2^{(i)}}, X_{D\cup J}))$ has a strictly positive Doob-Radon-Nikodym derivative w.r.t. $\mu_{B_1^{(i)}\cup C}$. Furthermore, the converse is also true for σ-finite reference measures $\mu_{B_1^{(i)}\cup C}$ by Corollary 2.3.20.

3. *Insertion/deletion of action: Assume:*

$$A \underset{G_{\mathrm{do}(I_B \cup D)}}{\overset{d}{\perp}} I_B \,|\, C \cup D.$$

For a fixed finite index set I consider subsets $B^{(i)} \subseteq B$, for $i \in I$, and pick for each $i \in I$ an arbitrary version of a conditional Markov kernel:

$$P(X_A |\, \mathrm{do}(X_{B^{(i)}}), X_C, \mathrm{do}(X_{D \cup J})):\ \mathcal{X}_{B \cup C \cup D \cup J} \to \mathcal{X}_{B^{(i)} \cup C \cup D \cup J} \to \mathcal{P}(\mathcal{X}_A),$$

of $P(X_A, X_C |\, \mathrm{do}(X_{B^{(i)}}, X_{D \cup J}))$ and assume the following absolute continuities:

$$\mu_C \ll P(X_C |\, \mathrm{do}(X_{B^{(i)}}, X_{D \cup J}))$$

for all $i \in I$. Then there exists a measurable μ_C-null set $N \subseteq \mathcal{X}_{B \cup C \cup D \cup J}$, such that all those conditional Markov kernels are equal on the complement N^{c}.

If we also assume the reverse absolute continuities for all $i \in I$:

$$P(X_C |\, \mathrm{do}(X_{B^{(i)}}, X_{D \cup J})) \ll \mu_C,$$

then all those conditional Markov kernels are versions of each other.[23]

The proof can be found in at the end of this section.
We now summarize on how to apply Theorem 5.1.2 more concretely as a corollary.

Corollary 5.1.3 (Almost-sure do-calculus—simplified)**.** *Consider an L-CBN:*

$$M = \left(G^+ = \left(J, (V, U), E^+\right), \left(P_v(X_v |\, \mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)}))\right)_{v \in V \cup U}\right).$$

Further assume that we have reference measures μ_v on $\mathcal{X}_v$ for every $v \in V$. We then put $\mu_F := \bigotimes_{v \in F} \mu_v$ for $F \subseteq V$. Let $A, B, C \subseteq V$ and $D \subseteq J \cup V$ be such that A, B, C, D are pairwise disjoint. Then we have the following 4 rules relating marginal conditional to marginal interventional Markov kernels:

1. *Insertion/deletion of observation, for $J \subseteq D$: Assume that we want to establish the a.s.-equality:*

$$P(X_A | X_B, X_C, \mathrm{do}(X_D)) = P(X_A | X_C, \mathrm{do}(X_D)) \qquad \mu_{B \cup C}\text{-a.s.},$$

then it is sufficient to assume/check the following d-separation and absolute continuities:

$$A \underset{G_{\mathrm{do}(D)}}{\overset{d}{\perp}} B \,|\, C \cup D, \qquad \mu_{B \cup C} \ll P(X_B, X_C |\, \mathrm{do}(X_D)) \ll \mu_{B \cup C}.$$

[23]Note that absolute continuities: $\mu_C \ll P(X_C |\, \mathrm{do}(X_{B^{(i)}}, X_{D \cup J})) \ll \mu_C$ hold if $P(X_C |\, \mathrm{do}(X_{B^{(i)}}, X_{D \cup J}))$ has a strictly positive Doob-Radon-Nikodym derivative w.r.t. μ_C. Furthermore, the converse is also true for σ-finite reference measures μ_C by Corollary 2.3.20.

2. <u>*Action/observation exchange*</u>, *for $J \subseteq D$: Assume that we want to establish the a.s.-equality:*

$$P(X_A|X_B, X_C, \mathrm{do}(X_D)) = P(X_A|\,\mathrm{do}(X_B), X_C, \mathrm{do}(X_D)) \qquad \mu_{B\cup C}\text{-a.s.},$$

then it is sufficient to assume/check the following d-separation and absolute continuities:

$$A \underset{G_{\mathrm{do}(I_B \cup D)}}{\overset{d}{\perp}} I_B \,|\, B \cup C \cup D, \qquad \mu_{B\cup C} \ll P(X_B, X_C|\,\mathrm{do}(X_D)) \ll \mu_{B\cup C},$$
$$\mu_C \ll P(X_C|\,\mathrm{do}(X_B, X_D)) \ll \mu_C.$$

3. <u>*Insertion/deletion of action*</u>, *for $J \subseteq D$: Assume that we want to establish the a.s.-equality:*

$$P(X_A|\,\mathrm{do}(X_B), X_C, \mathrm{do}(X_D)) = P(X_A|X_C, \mathrm{do}(X_D)) \qquad \mu_C\text{-a.s.},$$

then it is sufficient to assume/check the following d-separation and absolute continuities:

$$A \underset{G_{\mathrm{do}(I_B \cup D)}}{\overset{d}{\perp}} I_B \,|\, C \cup D, \qquad \mu_C \ll P(X_C|\,\mathrm{do}(X_B, X_D)) \ll \mu_C,$$
$$\mu_C \ll P(X_C|\,\mathrm{do}(X_D)) \ll \mu_C.$$

4. <u>*Deletion of input*</u>*: If*

$$A \underset{G_{\mathrm{do}(D)}}{\overset{d}{\perp}} J \,|\, C \cup D, \qquad \mu_C \ll P(X_C|\,\mathrm{do}(X_{D\cup J})) \ll \mu_C.$$

then there exists a Markov kernel $P(X_A|X_C, \mathrm{do}(X_D, \cancel{X_{J\setminus D}}))$ such that:

$$P(X_A|X_C, \mathrm{do}(X_D, \cancel{X_{J\setminus D}})) = P(X_A|X_C, \mathrm{do}(X_{D\cup J})) \qquad \mu_C\text{-a.s.}.$$

Note that the two-sided absolute continuities hold for σ-finite reference measures iff the indicated Markov kernel has a strictly positive Doob-Radon-Nikodym derivative w.r.t. the corresponding reference product measure by Corollary 2.3.20.

Proof. The proof follows directly from Theorem 5.1.2. For the last rule ('Deletion of input'), one can take the Markov kernel as

$$P(X_A|X_C, \mathrm{do}(X_D, \cancel{X_{J\setminus D}})) := Q(X_A|X_C, X_D)$$

where $Q(X_A|X_C, X_D)$ is defined in the proof of Proposition 5.1.8 point 3, for the special case $B = I_B = \emptyset$. The proof of Theorem 5.1.2 rule 3 then applies (as it doesn't depend crucially on the assumption $J \subseteq D$, or $B \neq \emptyset$), which shows the claim. $\square$

Remark 5.1.4. *We have made the following considerations in this particular formulation of the do-calculus.*

- *The additional assumptions $J \subseteq D$ in rule 1–3 were added because it is not clear whether (or how) a Markov kernel $P(X_A|X_C, \mathrm{do}(X_D))$ is defined for $J \not\subseteq D$. Rule 4 was added to make this possible.*

- *The assumption $J \subseteq D$ in rules 1–3 in Corollary 5.1.3 can be made without loss of generality and without restricting the applicability of the result. For example, for rule 1, if $\tilde{D}$ with $J \not\subseteq \tilde{D}$, then $A \perp^d_{G_{\mathrm{do}(\tilde{D})}} B|C \cup \tilde{D}$ implies $A \perp^d_{G_{\mathrm{do}(\tilde{D})}} B \cup J|C \cup \tilde{D}$, which implies $A \perp^d_{G_{\mathrm{do}(\tilde{D}\cup J)}} B|C \cup (\tilde{D} \cup J)$, and hence we can apply rule 1 to $D = \tilde{D} \cup J$ instead of $\tilde{D}$. Similar reasoning can be applied to show for rules 2 and 3 that the assumptions $J \subseteq D$ can be made without loss of generality.*

-

-

Remark 5.1.5. *Note that in rules 1–3 of Corollary 5.1.3 (in contrast to Proposition 5.1.8 and Theorem 5.1.2) we cannot easily formulate the independence of variables $X_{J\setminus D}$ in the presented way. This is the reason we decided to simplify the formulas in Corollary 5.1.3 by assuming that $J \subseteq D$ and thus $D \cup J = D$, which makes then the d-separation requirements weaker (due to extra conditioning on J). To compensate for the ensuing loss of generality, we have added the fourth rule. For the case where J does not fully lie in D one has several options:*

1. *use Proposition 5.1.8 or Theorem 5.1.2, or*
2. *use the global Markov property (Theorem 4.2.1) directly, or*
3. *combine rules 1–3 of Corollary 5.1.3 with rule 4 of Corollary 5.1.3.*

Proofs - Do-Calculus

Lemma 5.1.6. *For $B, C \subseteq V$ and $D \subseteq V \cup J$ with B, C, D pairwise disjoint, and a measurable subset $N \subseteq \mathcal{X}_{B\cup C\cup D\cup J}$ the following statements are equivalent:*

1. *N is a $P(X_B, X_C|\,\mathrm{do}(X_{I_B}, X_{D\cup J}))$-null set.*
2. *For every decomposition $B = B_1 \dot{\cup} B_2$ the set N is a $P(X_B, X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J}))$-null set.*
3. *For every decomposition $B = B_1 \dot{\cup} B_2$ the set N is a $P(X_{B_1}, X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J}))$-null set.*

Proof. Every value $x_{I_B} = (x_{I_v})_{v\in B} \in \mathcal{X}_{I_B}$ defines a decomposition $B = B_1 \dot\cup B_2$ via:

$$B_1 := \{v \in B \,|\, x_{I_v} = \star\}, \qquad B_2 := \{v \in B \,|\, x_{I_v} \in \mathcal{X}_v\}.$$

So running through all values $x_{I_B} \in \mathcal{X}_{I_B}$ is the same as running through all subsets $B_2 \subseteq B$ and all values $x_{B_2} \in \mathcal{X}_{B_2}$, while putting $x_{I_{B_1}} = \star$ for $B_1 = B \setminus B_2$. Furthermore, we have the following identities:

$$\begin{aligned}
&P((X_B, X_C) \in N_{x_{D\cup J}} |\, \mathrm{do}(X_{I_B} = x_{I_B}, X_{D\cup J} = x_{D\cup J}))\\
&= P((X_B, X_C) \in N_{x_{D\cup J}} |\, \mathrm{do}(X_{I_{B_1}} = \star, X_{I_{B_2}} = x_{B_2}, X_{D\cup J} = x_{D\cup J}))\\
&= P((X_B, X_C) \in N_{x_{D\cup J}} |\, \mathrm{do}(X_{B_2} = x_{B_2}, X_{D\cup J} = x_{D\cup J}))\\
&= (P(X_{B_1}, X_C |\, \mathrm{do}(X_{B_2} = x_{B_2}, X_{D\cup J} = x_{D\cup J})) \otimes \delta(X_{B_2}|X_{B_2} = x_{B_2}))\,(N_{x_{D\cup J}})\\
&= P((X_{B_1}, X_C) \in N_{(x_{B_2}, x_{D\cup J})} |\, \mathrm{do}(X_{B_2} = x_{B_2}, X_{D\cup J} = x_{D\cup J})).
\end{aligned}$$

So the first line vanishes for all values $x_{I_B} \in \mathcal{X}_{I_B}$ and $x_{D\cup J} \in \mathcal{X}_{D\cup J}$ if and only if any other line vanishes for all subsets $B_2 \subseteq B$ and all values $x_{B_2} \in \mathcal{X}_{B_2}$ and $x_{D\cup J} \in \mathcal{X}_{D\cup J}$. This shows the claim. □

Remark 5.1.7 (Null sets—again)**.** *In the following we will often make statements like: "The Markov kernel $K(X_A|X_B, X_C, X_D)$ is unique up to a measurable $K(X_B|X_C, X_D)$-null set in $\mathcal{X}_{B\cup C}$", (rather than in $\mathcal{X}_{B\cup C\cup D}$). This means that the corresponding null set N can be considered constant in $X_{D\setminus(B\cup C)}$, or, more precisely, that N is of the form:*

$$N = M \times \mathcal{X}_{D\setminus(B\cup C)} \subseteq \mathcal{X}_{B\cup C\cup D},$$

with $M \subseteq \mathcal{X}_{B\cup C}$.

Proposition 5.1.8 (Do-calculus—existence and uniqueness)**.** *Consider an L-CBN:*

$$M = \left(G^+ = \left(J, (V, U), E^+\right), \left(P_v(X_v |\, \mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)}))\right)_{v\in V\cup U}\right).$$

Let $A, B, C \subseteq V$ and $D \subseteq J \cup V$ be such that A, B, C, D are pairwise disjoint.
Then we have the following 3 rules relating marginal conditional to marginal interventional Markov kernels:

1. *Insertion/deletion of observation: If we have:*

$$A \underset{G_{\mathrm{do}(D)}}{\overset{d}{\perp}} B \,|\, C \cup D,$$

then there exists a Markov kernel:

$$P\left(X_A | \cancel{X_B}, X_C, \mathrm{do}(X_{D\cup \cancel{J}})\right) : \mathcal{X}_C \times \mathcal{X}_D \dashrightarrow \mathcal{X}_A,$$

that is a version of:

$$P\left(X_A | X_{B_2}, X_C, \mathrm{do}(X_{D\cup J})\right),$$

for every subset $B_2 \subseteq B$ simultaneously. Note that this Markov kernel is only dependent on x_C and x_D, and constant in $x_{J\setminus D}$.

Such a Markov kernel is unique up to a measurable $P(X_C|\,\mathrm{do}(X_{D\cup J}))$-null set in $\mathcal{X}_{C\cup D}$.

2. *Action/observation exchange: If we have:*

$$A \underset{G_{\mathrm{do}(I_B\cup D)}}{\overset{d}{\perp}} I_B \,|\, B\cup C\cup D,$$

then there exists a Markov kernel:

$$P\left(X_A|\cancel{\mathrm{do}}(X_B), X_C, \mathrm{do}(X_{D\cup\cancel{J}})\right) :\; \mathcal{X}_B\times\mathcal{X}_C\times\mathcal{X}_D \dashrightarrow \mathcal{X}_A,$$

that is a version of:

$$P\left(X_A|X_{B_1}, \mathrm{do}(X_{B_2}), X_C, \mathrm{do}(X_{D\cup J})\right),$$

for every decomposition: $B = B_1 \dot\cup B_2$, simultaneously.

Such a Markov kernel is unique up to a measurable $P(X_B, X_C|\,\mathrm{do}(X_{I_B}, X_{D\cup J}))$-null set $N\subseteq \mathcal{X}_{B\cup C\cup D}$, i.e. N is a $P(X_{B_1}, X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J}))$-null set for every decomposition $B = B_1\dot\cup B_2$ simultaneously.

3. *Insertion/deletion of action: If we have:*

$$A \underset{G_{\mathrm{do}(I_B\cup D)}}{\overset{d}{\perp}} I_B \,|\, C\cup D,$$

then there exists a Markov kernel:

$$P\left(X_A|\cancel{\mathrm{do}(X_B)}, X_C, \mathrm{do}(X_{D\cup\cancel{J}})\right) :\; \mathcal{X}_C\times\mathcal{X}_D \dashrightarrow \mathcal{X}_A,$$

that is a version of:

$$P\left(X_A|\,\mathrm{do}(X_{B_2}), X_C, \mathrm{do}(X_{D\cup J})\right)$$

for every subset $B_2 \subseteq B$ simultaneously. Note that this Markov kernel is only dependent on x_C and x_D, and constant in $x_{J\setminus D}$.

Such a Markov kernel is unique up to a measurable $P\left(X_C|\,\mathrm{do}(X_{I_B}, X_{D\cup J})\right)$-null set $N\subseteq\mathcal{X}_{C\cup D}$, i.e. N is a $P(X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J}))$-null set for every subset $B_2\subseteq B$ simultaneously.

Proof. We make use of the global Markov property (GMP), theorem 4.2.1.
1.) The assumption:

$$A \underset{G_{\mathrm{do}(D)}}{\overset{d}{\perp}} B \,|\, C\cup D,$$

implies the conditional independence by GMP 4.2.1:

$$X_A \underset{P(X_V|\,\mathrm{do}(X_{D\cup J}))}{\perp\!\!\!\perp} X_B \,|\, X_C, X_D.$$

So we get the following factorization, where we can omit the deterministic variables from X_D on the left of the conditioning lines:

$$P\left(X_A, X_B, X_C \middle| \operatorname{do}(X_{D\cup J})\right) = Q(X_A|X_C, X_D) \otimes P\left(X_B, X_C \middle| \operatorname{do}(X_{D\cup J})\right),$$

for some Markov kernel $Q(X_A|X_C, X_D)$. Here $Q(X_A|X_C, X_D)$ serves as a version of the conditional Markov kernel:

$$P\left(X_A|X_B, X_C, \operatorname{do}(X_{D\cup J})\right).$$

If we marginalize out X_{B_1} for any decomposition $B = B_1 \dot\cup B_2$ in the above factorization we also get:

$$P\left(X_A, X_C, X_{B_2} \middle| \operatorname{do}(X_{D\cup J})\right) = Q(X_A|X_C, X_D) \otimes P\left(X_C, X_{B_2} \middle| \operatorname{do}(X_{D\cup J})\right),$$

showing that $Q(X_A|X_C, X_D)$ is also a version of:

$$P\left(X_A|X_{B_2}, X_C, \operatorname{do}(X_{D\cup J})\right).$$

In particular, this holds for $B_2 = \emptyset$. This shows all the claimed properties for $Q(X_A|X_C, X_D)$.

Now consider another Markov kernel $K(X_A|X_C, X_D)$ and the measurable sets:

$$\begin{aligned}
\tilde{N} &:= \{x_{C\cup D} \in \mathcal{X}_{C\cup D} \,|\, Q(X_A|X_C = x_C, X_D = x_D) \neq K(X_A|X_C = x_C, X_D = x_D)\},\\
N &:= \{x_{C\cup D\cup J} \in \mathcal{X}_{C\cup D\cup J} \,|\, Q(X_A|X_C = x_C, X_D = x_D) \neq K(X_A|X_C = x_C, X_D = x_D)\}\\
&= \tilde{N} \times \mathcal{X}_{J\setminus D}.
\end{aligned}$$

If $K(X_A|X_C, X_D)$ is a version of:

$$P(X_A|X_{B_2}, X_C, \operatorname{do}(X_{D\cup J})),$$

for every subset $B_2 \subseteq B$ simultaneously, then this holds, in particular, for $B_2 = \emptyset$. Since conditional Markov kernels are essentially unique, by Theorem 2.4.16, we have that N is a $P(X_C|\operatorname{do}(X_{D\cup J}))$-null set.

2.) The assumption:

$$A \underset{G_{\operatorname{do}(I_B\cup D)}}{\overset{d}{\perp}} I_B \,|\, B \cup C \cup D,$$

implies the conditional independence by GMP 4.2.1:

$$X_A \underset{P(X_V|\operatorname{do}(X_{I_B}, X_{D\cup J}))}{\perp\!\!\!\perp} X_{I_B} \,|\, X_B, X_C, X_D.$$

So we have the following factorization:

$$P\left(X_A, X_B, X_C \middle| \operatorname{do}(X_{I_B}, X_{D\cup J})\right) = Q(X_A|X_B, X_C, X_D) \otimes P\left(X_B, X_C \middle| \operatorname{do}(X_{I_B}, X_{D\cup J})\right), \tag{21}$$

for some Markov kernel $Q(X_A|X_B, X_C, X_D)$, which serves as a version of the conditional Markov kernel:

$$P\left(X_A|X_B, X_C, \mathrm{do}(X_{I_B}, X_{D\cup J})\right),$$

and which is independent of X_{I_B}.

We first claim that for a Markov kernel $Q(X_A|X_B, X_C, X_D)$ the equation 21 is equivalent to the system of equations 22 indexed by subsets $B_2 \subseteq B$ and with $B_1 := B \setminus B_2$:

$$P\left(X_A, X_{B_1}, X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J})\right) = Q(X_A|X_B, X_C, X_D) \otimes P\left(X_{B_1}, X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J})\right). \tag{22}$$

Indeed, we can look at the different input values for $X_{I_B} = (X_{I_{B_1}}, X_{I_{B_2}})$ in equation 21. For B_1 we put: $X_{I_{B_1}} = \star = (\star)_{v\in B_1}$ and for B_2 we take values: $X_{I_{B_2}} = x_{B_2} \in \mathcal{X}_{B_2}$. This implies:

$$\begin{aligned}
&P\left(X_A, X_B, X_C|\,\mathrm{do}(X_{B_2} = x_{B_2}, X_{D\cup J})\right)\\
&= P\left(X_A, X_B, X_C|\,\mathrm{do}(X_{I_{B_1}} = \star, X_{I_{B_2}} = x_{B_2}, X_{D\cup J})\right)\\
&= Q(X_A|X_B, X_C, X_D) \otimes P\left(X_B, X_C|\,\mathrm{do}(X_{I_{B_1}} = \star, X_{I_{B_2}} = x_{B_2}, X_{D\cup J})\right),\\
&= Q(X_A|X_B, X_C, X_D) \otimes P\left(X_B, X_C|\,\mathrm{do}(X_{B_2} = x_{B_2}, X_{D\cup J})\right).
\end{aligned}$$

So we get the equations:

$$P\left(X_A, X_B, X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J})\right) = Q(X_A|X_B, X_C, X_D) \otimes P\left(X_B, X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J})\right),$$

where we can further marginalize out the deterministic X_{B_2}:

$$P\left(X_A, X_{B_1}, X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J})\right) = Q(X_A|X_B, X_C, X_D) \otimes P\left(X_{B_1}, X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J})\right).$$

Note that we can also go back by multiplying with $\delta(X_{B_2}|X_{B_2})$. This shows the intermediate claim.

The equation 22 already implies that $Q(X_A|X_B, X_C, X_D)$ is a version of the conditional Markov kernel:

$$P\left(X_A|X_{B_1}, X_C, \mathrm{do}(X_{B_2}, X_{D\cup J})\right), \tag{23}$$

for every decomposition: $B = B_1 \dot\cup B_2$ simultaneously.

Now consider another Markov kernel $K(X_A|X_B, X_C, X_D)$ and the measurable sets:

$$\begin{aligned}
\tilde N &:= \big\{x_{B\cup C\cup D} \in \mathcal{X}_{B\cup C\cup D} \,\big|\, Q(X_A|X_B = x_B, X_C = x_C, X_D = x_D)\\
&\qquad\qquad \neq K(X_A|X_B = x_B, X_C = x_C, X_D = x_D)\big\},\\
N &:= \tilde N \times \mathcal{X}_{J\setminus D}.
\end{aligned}$$

If $K(X_A|X_B, X_C, X_D)$ now is also a version of the conditional Markov kernel (23) for every decomposition $B = B_1 \dot\cup B_2$ simultaneously, then N is a $P(X_{B_1}, X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J}))$-null set for every decomposition $B = B_1 \dot\cup B_2$, because conditional Markov kernels are essentially unique, see Theorem 2.4.16. By Lemma 5.1.6 this statement is equivalent for

N to be a $P(X_B, X_C|\,\mathrm{do}(X_{I_B}, X_{D\cup J}))$-null set.

3.) The assumption:

$$A \underset{G_{\mathrm{do}(I_B\cup D)}}{\overset{d}{\perp}} I_B \,|\, C \cup D,$$

implies the conditional independence by GMP 4.2.1:

$$X_A \underset{P(X_V|\,\mathrm{do}(X_{I_B}, X_{D\cup J}))}{\perp\!\!\!\perp} X_{I_B} \,|\, X_C, X_D.$$

So we have the following factorization:

$$P\left(X_A, X_C|\,\mathrm{do}(X_{I_B}, X_{D\cup J})\right) = Q(X_A|X_C, X_D) \otimes P\left(X_C|\,\mathrm{do}(X_{I_B}, X_{D\cup J})\right), \tag{24}$$

for some Markov kernel $Q(X_A|X_C, X_D)$, which serves as a version of the conditional Markov kernel:

$$P\left(X_A|X_C, \mathrm{do}(X_{I_B}, X_{D\cup J})\right),$$

and which is independent of X_{I_B}.
We can now look at the different input values for any decomposition: $B = B_1 \dot\cup B_2$. For this we put: $X_{I_{B_1}} = \star = (\star)_{v\in B_1}$ and $X_{I_{B_2}} = x_{B_2} \in \mathcal{X}_{B_2}$. This implies:

$$\begin{aligned}
&P\left(X_A, X_C|\,\mathrm{do}(X_{B_2} = x_{B_2}, X_{D\cup J})\right)\\
&= P\left(X_A, X_C|\,\mathrm{do}(X_{I_{B_1}} = \star, X_{I_{B_2}} = x_{B_2}, X_{D\cup J})\right)\\
&= Q(X_A|X_C, X_D) \otimes P\left(X_C|\,\mathrm{do}(X_{I_{B_1}} = \star, X_{I_{B_2}} = x_{B_2}, X_{D\cup J})\right),\\
&= Q(X_A|X_C, X_D) \otimes P\left(X_C|\,\mathrm{do}(X_{B_2} = x_{B_2}, X_{D\cup J})\right).
\end{aligned}$$

So we get for every subset $B_2 \subseteq B$:

$$P\left(X_A, X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J})\right) = Q(X_A|X_C, X_D) \otimes P\left(X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J})\right), \tag{25}$$

which shows that $Q(X_A|X_C, X_D)$ is a version of the conditional Markov kernel:

$$P\left(X_A|X_C, \mathrm{do}(X_{B_2}, X_{D\cup J})\right), \tag{26}$$

for every subset $B_2 \subseteq B$ simultaneously.
Now consider another Markov kernel $K(X_A|X_C, X_D)$ and the measurable sets:

$$\begin{aligned}
\tilde N &:= \{x_{C\cup D} \in \mathcal{X}_{C\cup D} \,|\, Q(X_A|X_C = x_C, X_D = x_D) \neq K(X_A|X_C = x_C, X_D = x_D)\},\\
N_{B_2} &:= \mathcal{X}_{B_2} \times \tilde N \times \mathcal{X}_{J\setminus D}.
\end{aligned}$$

Now assume that $K(X_A|X_C, X_D)$ is a version of the conditional Markov kernel in (26) for every subset $B_2 \subseteq B$ simultaneously. Then for every subset $B_2 \subseteq B$ set N_{B_2} is a $P(X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J}))$-null set, because of the essential uniqueness of conditional

Markov kernels, see Theorem 2.4.16. More concretely, for $x_{B_2} \in \mathcal{X}_{B_2}$ and $x_{D\cup J} \in \mathcal{X}_{D\cup J}$ we get the equations:

$$\begin{aligned} 0 &= P(X_C \in (N_{B_2})_{(x_{B_2},x_{D\cup J})} |\, \mathrm{do}(X_{B_2} = x_{B_2}, X_{D\cup J} = x_{D\cup J})) \\ &= P(X_C \in \tilde{N}_{x_D} |\, \mathrm{do}(X_{B_2} = x_{B_2}, X_{D\cup J} = x_{D\cup J})) \\ &= P(X_C \in \tilde{N}_{x_D} |\, \mathrm{do}(X_{I_{B_1}} = \star, X_{I_{B_2}} = x_{B_2}, X_{D\cup J} = x_{D\cup J})) \\ &= P(X_C \in \tilde{N}_{x_D} |\, \mathrm{do}(X_{I_B} = (\star, x_{B_2}), X_{D\cup J} = x_{D\cup J})). \end{aligned}$$

Since we have this for all decompositions $B = B_1 \dot{\cup} B_2$ and all values $x_{B_2} \in \mathcal{X}_{B_2}$ we are running through all values $x_{I_B} \in \mathcal{X}_{I_B}$. This shows that $\tilde{N}$ is a $P(X_C |\, \mathrm{do}(X_{I_B}, X_{D\cup J}))$-null set in $\mathcal{X}_{C\cup D}$. □

Theorem 5.1.2. *[Almost-sure do-calculus—in detail] Consider an L-CBN:*

$$M = \left(G^+ = \left(J, (V, U), E^+\right), \left(P_v(X_v |\, \mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)}))\right)_{v\in V\cup U}\right).$$

Let $A, B, C \subseteq V$ and $D \subseteq J \cup V$ be such that A, B, C, D are pairwise disjoint. Further assume that we have reference measures μ_v on $\mathcal{X}_v$ for every $v \in V$ that are each equivalent to a probability measure (in terms of absolute continuity).[24] *We then put $\mu_F := \bigotimes_{v\in F} \mu_v$ for $F \subseteq V$.*

1. *Insertion/deletion of observation: Assume:*

$$A \underset{G_{\mathrm{do}(D)}}{\overset{d}{\perp}} B \,|\, C \cup D.$$

For a fixed finite index set I consider subsets $B^{(i)} \subseteq B$, for $i \in I$, and pick for each $i \in I$ an arbitrary version of a conditional Markov kernel:

$$P\left(X_A | X_{B^{(i)}}, X_C, \mathrm{do}(X_{D\cup J})\right) : \mathcal{X}_{B\cup C\cup D\cup J} \to \mathcal{X}_{B^{(i)}\cup C\cup D\cup J} \to \mathcal{P}(\mathcal{X}_A),$$

of $P(X_A, X_{B^{(i)}}, X_C |\, \mathrm{do}(X_{D\cup J}))$. Then there exists a measurable $P(X_B, X_C |\, \mathrm{do}(X_{D\cup J}))$-null set $N \subseteq \mathcal{X}_{B\cup C\cup D\cup J}$, such that all those Markov kernels are equal on the complement N^{c}.

Note that if $\mu_{B\cup C} \ll P(X_B, X_C |\, \mathrm{do}(X_{D\cup J}))$ then N is also a $\mu_{B\cup C}$-null set, i.e. for every $x_{D\cup J} \in \mathcal{X}_{D\cup J}$ we have: $\mu_{B\cup C}(N_{x_{D\cup J}}) = 0$.

If we also have the reverse $P(X_B, X_C |\, \mathrm{do}(X_{D\cup J})) \ll \mu_{B\cup C}$ then we can change the above conditional Markov kernels on a $\mu_{B\cup C}$-null set N while they remain versions of the corresponding conditional Markov kernel.[25]

[24] Recall the connection between absolute continuity and strictly positive densities in Corollary 2.3.20. All σ-finite measures satisfy this assumption.

[25] Note that the absolute continuities: $\mu_{B\cup C} \ll P(X_B, X_C |\, \mathrm{do}(X_{D\cup J})) \ll \mu_{B\cup C}$ hold if $P(X_B, X_C |\, \mathrm{do}(X_{D\cup J}))$ has a strictly positive Doob-Radon-Nikodym derivative w.r.t. $\mu_{B\cup C}$. Furthermore, the converse is also true for σ-finite reference measures $\mu_{B\cup C}$ by Corollary 2.3.20.

2. *Action/observation exchange:* Assume:

$$A \underset{G_{\mathrm{do}(I_B \cup D)}}{\overset{d}{\perp}} I_B \,|\, B \cup C \cup D.$$

For a fixed finite index set I consider decompositions $B = B_1^{(i)} \dot\cup B_2^{(i)}$, for $i \in I$, and pick for each $i \in I$ an arbitrary version of a conditional Markov kernel:

$$P(X_A | X_{B_1^{(i)}}, \mathrm{do}(X_{B_2^{(i)}}), X_C, \mathrm{do}(X_{D \cup J})) : \mathcal{X}_{B \cup C \cup D \cup J} \to \mathcal{P}(\mathcal{X}_A),$$

of $P(X_A, X_{B_1^{(i)}}, X_C |\, \mathrm{do}(X_{B_2^{(i)}}, X_{D\cup J})) \otimes \mu_{B_2^{(i)}}$ and assume the following absolute continuities:

$$\mu_{B \cup C} \ll P(X_{B_1^{(i)}}, X_C |\, \mathrm{do}(X_{B_2^{(i)}}, X_{D \cup J})) \otimes \mu_{B_2^{(i)}}$$

for all $i \in I$.[26] *Then there exists a measurable $\mu_{B\cup C}$-null set $N \subseteq \mathcal{X}_{B \cup C \cup D \cup J}$, such that all those conditional Markov kernels are equal on the complement N^{c}.*

If we also assume the reverse absolute continuities for all $i \in I$:

$$P(X_{B_1^{(i)}}, X_C |\, \mathrm{do}(X_{B_2^{(i)}}, X_{D \cup J})) \otimes \mu_{B_2^{(i)}} \ll \mu_{B \cup C},$$

then all those conditional Markov kernels are versions of each other.[27]

3. *Insertion/deletion of action:* Assume:

$$A \underset{G_{\mathrm{do}(I_B \cup D)}}{\overset{d}{\perp}} I_B \,|\, C \cup D.$$

For a fixed finite index set I consider subsets $B^{(i)} \subseteq B$, for $i \in I$, and pick for each $i \in I$ an arbitrary version of a conditional Markov kernel:

$$P(X_A |\, \mathrm{do}(X_{B^{(i)}}), X_C, \mathrm{do}(X_{D \cup J})) : \mathcal{X}_{B \cup C \cup D \cup J} \to \mathcal{X}_{B^{(i)} \cup C \cup D \cup J} \to \mathcal{P}(\mathcal{X}_A),$$

[26] If you instead expected to pick for each $i \in I$ an arbitrary version of a conditional Markov kernel:

$$P(X_A | X_{B_1^{(i)}}, \mathrm{do}(X_{B_2^{(i)}}), X_C, \mathrm{do}(X_{D \cup J})) : \mathcal{X}_{B \cup C \cup D \cup J} \to \mathcal{P}(\mathcal{X}_A),$$

of $P(X_A, X_{B_1^{(i)}}, X_C |\, \mathrm{do}(X_{B_2^{(i)}}, X_{D \cup J}))$ and to assume the absolute continuities

$$\mu_{B_1^{(i)} \cup C} \ll P(X_{B_1^{(i)}}, X_C |\, \mathrm{do}(X_{B_2^{(i)}}, X_{D \cup J})) \ll \mu_{B_1^{(i)} \cup C}$$

for all $i \in I$: that would lead to a similar, but slightly weaker statement.

[27] Note that the absolute continuities: $\mu_{B \cup C} \ll P(X_{B_1^{(i)}}, X_C |\, \mathrm{do}(X_{B_2^{(i)}}, X_{D \cup J})) \otimes \mu_{B_2^{(i)}} \ll \mu_{B \cup C}$ hold if the absolute continuities: $\mu_{B_1^{(i)} \cup C} \ll P(X_{B_1^{(i)}}, X_C |\, \mathrm{do}(X_{B_2^{(i)}}, X_{D \cup J})) \ll \mu_{B_1^{(i)} \cup C}$ hold, which hold if $P(X_{B_1^{(i)}}, X_C |\, \mathrm{do}(X_{B_2^{(i)}}, X_{D \cup J}))$ has a strictly positive Doob-Radon-Nikodym derivative w.r.t. $\mu_{B_1^{(i)} \cup C}$. Furthermore, the converse is also true for σ-finite reference measures $\mu_{B_1^{(i)} \cup C}$ by Corollary 2.3.20.

of $P(X_A, X_C|\,\mathrm{do}(X_{B^{(i)}}, X_{D\cup J}))$ and assume the following absolute continuities:

$$\mu_C \ll P(X_C|\,\mathrm{do}(X_{B^{(i)}}, X_{D\cup J}))$$

for all $i \in I$. Then there exists a measurable μ_C-null set $N \subseteq \mathcal{X}_{B\cup C\cup D\cup J}$, such that all those conditional Markov kernels are equal on the complement N^{c}.

If we also assume the reverse absolute continuities for all $i \in I$:

$$P(X_C|\,\mathrm{do}(X_{B^{(i)}}, X_{D\cup J})) \ll \mu_C,$$

then all those conditional Markov kernels are versions of each other.[28]

Proof. W.l.o.g. we can assume all μ_v to be probability measures.

1.) Let $Q(X_A|X_C, X_D)$ be the Markov kernel from Proposition 5.1.8 point 1. Recall that conditional Markov kernels are essentially unique by Theorem 2.4.16. This shows that the set:

$$\begin{aligned}\tilde{N}^{(i)} := \big\{x_{B^{(i)}\cup C\cup D\cup J} \in \mathcal{X}_{B^{(i)}\cup C\cup D\cup J} \,\big|\, & Q(X_A|X_C = x_C, X_D = x_D) \\ & \neq P(X_A|X_{B^{(i)}} = x_{B^{(i)}}, X_C = x_C, \mathrm{do}(X_{D\cup J} = x_{D\cup J}))\big\},\end{aligned}$$

is a (measurable) $P(X_{B^{(i)}}, X_C|\,\mathrm{do}(X_{D\cup J}))$-null set. So the lifted set:

$$\begin{aligned}N^{(i)} := \big\{x_{B\cup C\cup D\cup J} \in \mathcal{X}_{B\cup C\cup D\cup J} \,\big|\, & Q(X_A|X_C = x_C, X_D = x_D) \\ & \neq P(X_A|X_{B^{(i)}} = x_{B^{(i)}}, X_C = x_C, \mathrm{do}(X_{D\cup J} = x_{D\cup J}))\big\},\end{aligned}$$

is then a (measurable) $P(X_B, X_C|\,\mathrm{do}(X_{D\cup J}))$-null set. Then also the finite union:

$$N := \bigcup_{i\in I} N^{(i)} \subseteq \mathcal{X}_{B\cup C\cup D\cup J},$$

is a (measurable) $P(X_B, X_C|\,\mathrm{do}(X_{D\cup J}))$-null set as well. Note that on the complement N^{c} all Markov kernels agree with $Q(X_A|X_C, X_D)$ and are thus all equal on N^{c}.

2.) Consider the Markov kernel $Q(X_A|X_B, X_C, X_D)$ from Proposition 5.1.8 point 2 and for $i \in I$ the measurable set:

$$\begin{aligned}N^{(i)} := \big\{x_{B\cup C\cup D\cup J} \in \mathcal{X}_{B\cup C\cup D\cup J} \,\big|\, & Q(X_A|X_B = x_B, X_C = x_C, X_D = x_D) \\ \neq P(X_A|X_{B_1^{(i)}} = x_{B_1^{(i)}}, \mathrm{do}(X_{B_2^{(i)}} = x_{B_2^{(i)}}), & X_C = x_C, \mathrm{do}(X_{D\cup J} = x_{D\cup J}))\big\}.\end{aligned}$$

Again, by the essential uniqueness of conditional Markov kernels, Theorem 2.4.16, the set $N^{(i)}$ is a $P(X_{B_1^{(i)}}, X_C|\,\mathrm{do}(X_{B_2^{(i)}}, X_{D\cup J})) \otimes \mu_{B_2^{(i)}}$-null set. The absolute continuity:

$$\mu_{B\cup C} \ll P(X_{B_1^{(i)}}, X_C|\,\mathrm{do}(X_{B_2^{(i)}}, X_{D\cup J})) \otimes \mu_{B_2^{(i)}},$$

[28] Note that absolute continuities: $\mu_C \ll P(X_C|\,\mathrm{do}(X_{B^{(i)}}, X_{D\cup J})) \ll \mu_C$ hold if $P(X_C|\,\mathrm{do}(X_{B^{(i)}}, X_{D\cup J}))$ has a strictly positive Doob-Radon-Nikodym derivative w.r.t. μ_C. Furthermore, the converse is also true for σ-finite reference measures μ_C by Corollary 2.3.20.

then renders $N^{(i)}$ a $\mu_{B\cup C}$-null set. This shows that the finite union:

$$N := \bigcup_{i\in I} N^{(i)},$$

is a $\mu_{B\cup C}$-null set as well. Again, note that on the complement $N^\complement$ all Markov kernels agree with $Q(X_A|X_B, X_C, X_D)$ and are thus all equal on $N^\complement$.

3.) Consider the Markov kernel $Q(X_A|X_C, X_D)$ from Proposition 5.1.8 point 3 and for $i \in I$ the measurable set:

$$\begin{aligned}\tilde{N}^{(i)} := \big\{x_{B^{(i)}\cup C\cup D\cup J} \in \mathcal{X}_{B^{(i)}\cup C\cup D\cup J} \,\big|\, & Q(X_A|X_C = x_C, X_D = x_D) \\ & \neq P(X_A|\operatorname{do}(X_{B^{(i)}} = x_{B^{(i)}}), X_C = x_C, \operatorname{do}(X_{D\cup J} = x_{D\cup J}))\big\}.\end{aligned}$$

Again, by the essential uniqueness of conditional Markov kernels, Theorem 2.4.16, the set $\tilde{N}^{(i)}$ is a $P(X_C|\operatorname{do}(X_{B^{(i)}}, X_{D\cup J}))$-null set. By the absolute continuity $\mu_C \ll P(X_C|\operatorname{do}(X_{B^{(i)}}, X_{D\cup J}))$ we get that $\tilde{N}^{(i)}$ is a μ_C-null set. This shows that the measurable set:

$$\begin{aligned}N^{(i)} := \big\{x_{B\cup C\cup D\cup J} \in \mathcal{X}_{B\cup C\cup D\cup J} \,\big|\, & Q(X_A|X_C = x_C, X_D = x_D) \\ & \neq P(X_A|\operatorname{do}(X_{B^{(i)}} = x_{B^{(i)}}), X_C = x_C, \operatorname{do}(X_{D\cup J} = x_{D\cup J}))\big\}.\end{aligned}$$

is a μ_C-null set as well. Then the finite union:

$$N := \bigcup_{i\in I} N^{(i)},$$

is also a μ_C-null set. Again, note that on the complement $N^\complement$ all Markov kernels agree with $Q(X_A|X_C, X_D)$ and are thus all equal on $N^\complement$. □

5.2. Adjustment Criteria and Formulae

Motivation 5.2.1. *Consider an L-CBN:*

$$M = \left(G^+ = \left(J, (V, U), E^+\right), \left(P_v(X_v|\operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)}))\right)_{v\in V\cup U}\right).$$

*For simplicity assume that there are no input variables, i.e. $J = \emptyset$. Then the joint distribution is "*do*-free" and given as:*

$$P(X_V, X_U) = \bigotimes_{v\in U\cup V} P_v\left(X_v|\operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)})\right),$$

with observational distribution as its marginal: $P(X_V)$.
We also have all the interventional distributions for $W \subseteq V$:

$$P(X_{V\setminus W}, X_U|\operatorname{do}(X_W)) = \bigotimes_{v\in U\cup V\setminus W} P_v\left(X_v|\operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)})\right),$$

with marginals: $P(X_{V\setminus W}|\,\mathrm{do}(X_W))$.
If we wanted to learn the distribution $P(X_V)$ *we could do an observational study and apply the usual statistical or machine learning techniques. If, in contrast, we wanted to learn interventional distributions:* $P(X_{V\setminus W}|\,\mathrm{do}(X_W))$ *from data (e.g. whether vaccination makes people immune to a disease), we typically would need to perform an interventional study where we intervene on the variables* X_W *and set them to different values. This usually requires expensive, time-consuming randomized control trials with an own group for each possible value of* X_W.
If we assume that we know the causal graph G^+ *or* G *we could try to leverage the rules of do-calculus in a clever way and might be able to go from expressions involving* $\mathrm{do}(W)$ *to expressions only involving* $\mathrm{do}(D)$ *for a (much) smaller subset* $D \subseteq W$*, ideally* $D = \emptyset$*. Practically this would mean that we would need a much smaller randomized control trial and save time and resources.*
For example, if we have the graph only involving the edge: $v_1 \longrightarrow v_2$ *we have that:*

$$P(X_2|\,\mathrm{do}(X_1)) = P(X_2|X_1),$$

which can be estimated using observational data only, e.g. via supervised learning.
So the question of identifiability is now: Assuming that the causal graph is known, under which circumstances is a causal effect $P(X_A|\,\mathrm{do}(X_B))$ *already determined by the observational distribution* $P(X_V)$*? When can causal effects be identified via distributions that have less interventions in them?*

Notation 5.2.2. *Consider an L-CBN:*

$$M = \left(G^+ = \left(J, (V,U), E^+\right), \left(P_v(X_v|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)}))\right)_{v \in V \cup U}\right).$$

We are interested in estimating the conditional causal effect:

$$P(X_A|X_C, \mathrm{do}(X_B, X_D)),$$

but we only have data from:

$$P(X_V|X_C, \mathrm{do}(X_D)).$$

We will consider the following disjoint subsets of variables:

1. $A \subseteq V$*: the outcome variables of interest;*
2. $B \subseteq V$*: the treatment or intervention variables;*
3. $C \subseteq V$*: general conditional (context) variables under which the data was collected;*
4. $D \subseteq J \cup V$*: general interventional (context) variables that were set by the experimenter (assuming w.l.o.g.* $J \subseteq D$*);*
5. $F_0 \subseteq V$*: core adjustment variables, i.e. features that were measured;*
6. $F_1 \subseteq V$*: additional measured adjustment variables;*

7. $H \subseteq U$: additional unobserved variables;

and write $F := F_0 \cup F_1$.

Theorem 5.2.3 (General adjustment formula)**.** *Consider an L-CBN:*

$$M = \left(G^+ = \left(J, (V, U), E^+\right), \left(P_v(X_v | \operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)}))\right)_{v \in V \cup U}\right).$$

Assume that all the following conditions hold in the graphs $G^+_{\operatorname{do}(I_B, D)}$:

$$(F_0 \cup H) \underset{G^+_{\operatorname{do}(I_B,D)}}{\overset{d}{\perp}} I_B \,|(C \cup D), \tag{27}$$

$$A \underset{G^+_{\operatorname{do}(I_B,D)}}{\overset{d}{\perp}} (F_1 \cup I_B) \,|(B \cup F_0 \cup H \cup C \cup D), \tag{28}$$

$$H \underset{G^+_{\operatorname{do}(I_B,D)}}{\overset{d}{\perp}} B \,|(F \cup C \cup I_B \cup D). \tag{29}$$

Further assume that we have reference measures μ_v on $\mathcal{X}_v$, $v \in V \cup H$, such that:

$$\mu_{B \cup C \cup F \cup H} \ll P(X_B, X_C, X_F, X_H | \operatorname{do}(X_D)) \ll \mu_{B \cup C \cup F \cup H},$$
$$\mu_{C \cup F \cup H} \ll P(X_C, X_F, X_H | \operatorname{do}(X_B, X_D)) \ll \mu_{C \cup F \cup H}.$$

Then we have the adjustment formula:

$$P(X_A | X_C, \operatorname{do}(X_B, X_D)) = P(X_A | X_B, X_C, X_F, \operatorname{do}(X_D)) \circ P(X_F | X_C, \operatorname{do}(X_D)) \qquad \mu_{B \cup C}\text{-a.s.}$$

Proof. With help of Corollary 5.1.3 (2nd rule) we can establish the a.s.-equality:

$$P(X_A | X_{F_0}, X_H, X_C, X_B, \operatorname{do}(X_D)) = P(X_A | X_{F_0}, X_H, X_C, \operatorname{do}(X_B, X_D)) \quad \mu_{F_0 \cup H \cup C \cup B}\text{-a.s.}, \tag{30}$$

using the assumptions (implied by eq. 28):

$$A \underset{G^+_{\operatorname{do}(I_B,D)}}{\overset{d}{\perp}} I_B \,|(B \cup F_0 \cup H \cup C \cup D),$$
$$\mu_{F_0 \cup H \cup C \cup B} \ll P(X_{F_0}, X_H, X_C, X_B | \operatorname{do}(X_D)) \ll \mu_{F_0 \cup H \cup C \cup B},$$
$$\mu_{F_0 \cup H \cup C} \ll P(X_{F_0}, X_H, X_C | \operatorname{do}(X_B, X_D)) \ll \mu_{F_0 \cup H \cup C}.$$

With help of Corollary 5.1.3 (3rd rule) we can establish the a.s.-equality:

$$P(X_{F_0}, X_H | X_C, \operatorname{do}(X_B, X_D)) = P(X_{F_0}, X_H | X_C, \operatorname{do}(X_D)) \qquad \mu_C\text{-a.s.}, \tag{31}$$

using the assumptions (implied by eq. 27):

$$(F_0 \cup H) \underset{G^+_{\operatorname{do}(I_B,D)}}{\overset{d}{\perp}} I_B \,|(C \cup D),$$
$$\mu_C \ll P(X_C | \operatorname{do}(X_B, X_D)) \ll \mu_C,$$
$$\mu_C \ll P(X_C | \operatorname{do}(X_D)) \ll \mu_C.$$

With help of Corollary 5.1.3 (1st rule) we can establish the a.s.-equality:

$$P(X_A|X_{F_0}, X_H, X_C, X_B, \mathrm{do}(X_D)) = P(X_A|X_{F_0}, X_{F_1}, X_H, X_C, X_B, \mathrm{do}(X_D)) \qquad (32)$$
$$\mu_{F_0 \cup F_1 \cup H \cup C \cup B}\text{-a.s.},$$

using the assumptions (implied by eq. 28):

$$A \underset{G^+_{\mathrm{do}(D)}}{\overset{d}{\perp}} F_1 \,|(B \cup F_0 \cup H \cup C \cup D),$$
$$\mu_{F_0 \cup F_1 \cup H \cup C \cup B} \ll P(X_{F_0}, X_{F_1}, X_H, X_C, X_B|\,\mathrm{do}(X_D)) \ll \mu_{F_0 \cup F_1 \cup H \cup C \cup B}.$$

With help of Corollary 5.1.3 (1st rule) we can establish the a.s.-equality:

$$P(X_H|X_F, X_C, \mathrm{do}(X_D)) = P(X_H|X_F, X_C, X_B, \mathrm{do}(X_D)) \qquad \mu_{F \cup C \cup B}\text{-a.s.}, \qquad (33)$$

using the assumptions (implied by eq. 29):

$$H \underset{G^+_{\mathrm{do}(D)}}{\overset{d}{\perp}} B \,|(F \cup C \cup D),$$
$$\mu_{F \cup C \cup B} \ll P(X_F, X_C, X_B|\,\mathrm{do}(X_D)) \ll \mu_{F \cup C \cup B}.$$

These a.s.-equations together with the chain rule give us the following $\mu_{B \cup C}$-a.s.-equation:

$$\begin{aligned}
&P(X_A|X_C, \mathrm{do}(X_B, X_D)) \\
&= P(X_A|X_{F_0}, X_H, X_C, \mathrm{do}(X_B, X_D)) \circ P(X_{F_0}, X_H|X_C, \mathrm{do}(X_B, X_D)) \\
&\overset{30}{=} P(X_A|X_{F_0}, X_H, X_C, X_B, \mathrm{do}(X_D)) \circ P(X_{F_0}, X_H|X_C, \mathrm{do}(X_B, X_D)) \\
&\overset{31}{=} P(X_A|X_{F_0}, X_H, X_C, X_B, \mathrm{do}(X_D)) \circ P(X_{F_0}, X_H|X_C, \mathrm{do}(X_D)) \\
&= P(X_A|X_{F_0}, X_H, X_C, X_B, \mathrm{do}(X_D)) \circ P(X_{F_0}, X_{F_1}, X_H|X_C, \mathrm{do}(X_D)) \\
&\overset{32}{=} P(X_A|X_{F_0}, X_{F_1}, X_H, X_C, X_B, \mathrm{do}(X_D)) \circ P(X_{F_0}, X_{F_1}, X_H|X_C, \mathrm{do}(X_D)) \\
&= P(X_A|X_F, X_H, X_C, X_B, \mathrm{do}(X_D)) \circ P(X_F, X_H|X_C, \mathrm{do}(X_D)) \\
&= P(X_A|X_F, X_H, X_C, X_B, \mathrm{do}(X_D)) \circ (P(X_H|X_F, X_C, \mathrm{do}(X_D)) \otimes P(X_F|X_C, \mathrm{do}(X_D))) \\
&\overset{33}{=} P(X_A|X_F, X_H, X_C, X_B, \mathrm{do}(X_D)) \circ (P(X_H|X_F, X_C, X_B, \mathrm{do}(X_D)) \otimes P(X_F|X_C, \mathrm{do}(X_D))) \\
&= P(X_A|X_F, X_C, X_B, \mathrm{do}(X_D)) \circ P(X_F|X_C, \mathrm{do}(X_D)).
\end{aligned}$$

Note that the disintegration:

$$P(X_F, X_H|X_C, \mathrm{do}(X_D)) = P(X_H|X_F, X_C, \mathrm{do}(X_D)) \otimes P(X_F|X_C, \mathrm{do}(X_D))$$

holds (only) $P(X_C|\,\mathrm{do}(X_D))$-a.s., as for the conditional $P(X_H|X_F, X_C, \mathrm{do}(X_D))$ we have the ambiguity if it is considered a conditional of $P(X_F, X_H|X_C, \mathrm{do}(X_D))$, for which then we have "sure" equality, or, if it is considered a conditional of $P(X_F, X_H, X_C|\,\mathrm{do}(X_D))$, for which then we have only the above "almost-sure" equality. Further note, that by the assumption on the reference measures, the above equality then also holds μ_C-a.s. $\square$

Corollary 5.2.4 (Conditional interventional backdoor covariate adjustment formula). *Consider an L-CBN:*

$$M = \left(G^+ = \left(J, (V, U), E^+\right), \left(P_v(X_v|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)})) \right)_{v \in V \cup U} \right).$$

Assume that the ***conditional interventional backdoor criterion*** *in the graphs $G_{\mathrm{do}(I_B,D)}$ holds:*

1. $F \underset{G_{\mathrm{do}(I_B,D)}}{\overset{d}{\perp}} I_B \,|(C \cup D)$, *and:*
2. $A \underset{G_{\mathrm{do}(I_B,D)}}{\overset{d}{\perp}} I_B \,|(B \cup F \cup C \cup D)$.

Further assume the following absolute continuities:

$$\begin{aligned} \mu_{B \cup C \cup F} \ll P(X_B, X_C, X_F|\,\mathrm{do}(X_D)) \ll \mu_{B \cup C \cup F}, \\ \mu_{C \cup F} \ll P(X_C, X_F|\,\mathrm{do}(X_B, X_D)) \ll \mu_{C \cup F}. \end{aligned}$$

Then we have the adjustment formula:

$$P(X_A|X_C, \mathrm{do}(X_B, X_D)) = P(X_A|X_B, X_C, X_F, \mathrm{do}(X_D)) \circ P(X_F|X_C, \mathrm{do}(X_D)) \qquad \mu_{B \cup C}\text{-a.s.}$$

Proof. It follows by the same arguments as in Theorem 5.2.3 with $F_1 = H = \emptyset$. $\square$

Without the conditioning set, i.e. $C = \emptyset$, and direct careful analysis we get a version with slightly weaker positivity assumptions:

Corollary 5.2.5 (Interventional backdoor covariate adjustment formula). *Consider an L-CBN:*

$$M = \left(G^+ = \left(J, (V, U), E^+\right), \left(P_v(X_v|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)})) \right)_{v \in V \cup U} \right).$$

Assume that the ***interventional backdoor criterion*** *in the graphs $G_{\mathrm{do}(I_B,D)}$ holds:*

1. $F \underset{G_{\mathrm{do}(I_B,D)}}{\overset{d}{\perp}} I_B \,|\, D$, *and:*
2. $A \underset{G_{\mathrm{do}(I_B,D)}}{\overset{d}{\perp}} I_B \,|(B \cup F \cup D)$.

Further assume the following absolute continuity:

$$P(X_F|\,\mathrm{do}(X_D)) \otimes P(X_B|\,\mathrm{do}(X_D)) \ll P(X_F, X_B|\,\mathrm{do}(X_D)).$$

Then we have the adjustment formulas:

$$\begin{aligned} P(X_A, X_F|\,\mathrm{do}(X_B, X_D)) &= P(X_A|X_F, X_B, \mathrm{do}(X_D)) \otimes P(X_F|\,\mathrm{do}(X_D)) \quad P(X_B|\,\mathrm{do}(X_D))\text{-a.s.}, \\ P(X_A|\,\mathrm{do}(X_B, X_D)) &= P(X_A|X_F, X_B, \mathrm{do}(X_D)) \circ P(X_F|\,\mathrm{do}(X_D)) \quad P(X_B|\,\mathrm{do}(X_D))\text{-a.s.} \end{aligned}$$

Proof. First, note that "$\mathrm{do}(X_D)$" appears in every Markov kernel in the above formulas. So for readability, we will drop it in the following everywhere.

By the first d-separation assumption we see by Proposition 5.1.8 (rule 3) that we have the "sure" equality: $P(X_F|\,\mathrm{do}(X_B)) = P(X_F)$. By the second d-separation assumption we see by Proposition 5.1.8 (rule 2) that we have a Markov kernel $P(X_A|X_F, \mathrm{do}(X_B))$ that is also a version of $P(X_A|X_F, X_B)$. So any version of $P(X_A|X_F, X_B)$ can be changed on a $P(X_F, X_B)$-null set $N \subseteq \mathcal{X}_B \times \mathcal{X}_F$ to get $P(X_A|X_F, \mathrm{do}(X_B))$. The absolute continuity assumption implies that N is also a $P(X_F) \otimes P(X_B)$-null set. This implies that we have the equations of Markov kernels:

$$\begin{aligned} P(X_A|X_F, X_B) \otimes P(X_F) \otimes P(X_B) &= P(X_A|X_F, \mathrm{do}(X_B)) \otimes P(X_F) \otimes P(X_B) \\ &= P(X_A|X_F, \mathrm{do}(X_B)) \otimes P(X_F|\,\mathrm{do}(X_B)) \otimes P(X_B) \\ &= P(X_A, X_F|\,\mathrm{do}(X_B)) \otimes P(X_B). \end{aligned}$$

By the essential uniqueness of conditional Markov kernels we get that:

$$P(X_A, X_F|\,\mathrm{do}(X_B)) = P(X_A|X_F, X_B) \otimes P(X_F) \qquad P(X_B)\text{-a.s.}$$

Marginalizing out X_F on both sides gives us the remaining claim. □

We can now further specialize to the case with $C = D = J = \emptyset$ and immediately get:

Corollary 5.2.6 (Backdoor covariate adjustment). *Assume that the* ***backdoor criterion*** *holds:*

1. $F \underset{G_{\mathrm{do}(I_B)}}{\overset{d}{\perp}} I_B$, *and:*

2. $A \underset{G_{\mathrm{do}(I_B)}}{\overset{d}{\perp}} I_B \,|(B \cup F)$.

Further assume the following absolute continuity:

$$P(X_F) \otimes P(X_B) \ll P(X_F, X_B).$$

Then we have the adjustment formulas:

$$\begin{aligned} P(X_A, X_F|\,\mathrm{do}(X_B)) &= P(X_A|X_F, X_B) \otimes P(X_F) \qquad && P(X_B)\text{-a.s.}, \\ P(X_A|\,\mathrm{do}(X_B)) &= P(X_A|X_F, X_B) \circ P(X_F) && P(X_B)\text{-a.s.} \end{aligned}$$

Remark 5.2.7. *An example how the adjustment formula may fail if the strict positivity assumptions are not met is provided in Example 5.3.29.*

5.3. The ID-Algorithm

Consider an L-CBN:

$$M = \left(G^+ = \left(J, (V, U), E^+\right), \left(P_v(X_v | \operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)})) \right)_{v \in V \cup U} \right),$$

with observable CADMG G. For disjoint subsets $A, B, C \subseteq V$ we want to infer the conditional interventional distribution $P(X_A | X_B, \operatorname{do}(X_{J \cup C}))$ in terms of (repeated products and) conditional marginals of the observable Markov kernel $P(X_V | \operatorname{do}(X_J))$ and knowledge of G. In this subsection we will restrict ourselves to the case $P(X_A | \operatorname{do}(X_{J \cup C}))$ with no conditioning and present the *ID-algorithm*, which can tell if this is possible or not (in precise terms), and if so, provides us with a formula to do so. We will start this subsection with a series of necessary definitions, notations and lemmata. The main references are [Pea09, GP95, Tia02, TP02, Tia04, SP06b, HV06, HV08, RERS23, FM20].

5.3.1. Core Definitions and Notations

Definition 5.3.1 (Identifiability of interventional distributions/Markov kernels)**.** *Let $G = (J, V, E, L)$ be a CADMG and $B \subseteq V$ and $C \subseteq J \cup V$ disjoint subsets. We say that the interventional distribution/Markov kernel of C onto B is **identifiable** from G, or, more in generic symbols, that $P(X_B | \operatorname{do}(X_{J \cup C}))$ is **identifiable** from $P(X_V | \operatorname{do}(X_J))$ (and G), if for every two L-CBNs M_1 and M_2 with the same:*

1. *observable CADMG $G_1 = G_2 = G$, and:*
2. *underlying spaces $\mathcal{X}_{1,v} = \mathcal{X}_{2,v} =: \mathcal{X}_v$ for $v \in J \cup V$, and:*
3. *observable Markov kernels $P_1(X_V | \operatorname{do}(X_J)) = P_2(X_V | \operatorname{do}(X_J))$,*

we also have the equality of the interventional Markov kernels:

$$P_1(X_B | \operatorname{do}(X_{J \cup C})) = P_2(X_B | \operatorname{do}(X_{J \cup C})).$$

Sometimes we further restrict the class of CBNs to define/achieve identifiability, e.g. by adding "for linear Gaussian CBNs" or "for discrete CBNs with strictly positive mass functions", etc., and then only require M_1 and M_2 to come from such classes.

In the following we introduce a somewhat more vague, but constructive, notion of identifiability, which we coin *trackability* that allows us to follow certain marginalization, conditioning and multiplication steps to arrive at the wanted interventional Markov kernel.

Definition 5.3.2 (Trackability (up to specifications))**.** *Let $G = (J, V, E, L)$ be a CADMG and $B \subseteq V$ and $C \subseteq J \cup V$ disjoint subsets.*

1. *We say that the interventional distribution/Markov kernel of C onto B is **trackable** from G, or simply that $P(X_B | \operatorname{do}(X_{J \cup C}))$ is **trackable** from $P(X_V | \operatorname{do}(X_J))$*

(and G), if there exists a finite sequence of operations, only involving marginalization, conditioning and multiplication of Markov kernels, applied to previously determined Markov kernels, starting from $P(X_V|\,\mathrm{do}(X_J))$, with predetermined target sets $T_n \subseteq G$, indicating on which variable the operation is applied to, such that for every L-CBNs M with observable CADMG G we can compute $P(X_B|\,\mathrm{do}(X_{J\cup C}))$ from $P(X_V|\,\mathrm{do}(X_J))$ when we follow the above sequence of operations (and this should work no matter which version of conditional Markov kernels were used).

2. *We say that $P(X_B|\,\mathrm{do}(X_{J\cup C}))$ is **trackable up to specifications** "xyz" from $P(X_V|\,\mathrm{do}(X_J))$ (and G), if the same as above holds true, but whenever we condition, which leads to a Markov kernel only up so some null sets, we use the specifications "xyz" to pick a certain version of conditional Markov kernel at each step such that following the pre-specified operations leads to $P(X_B|\,\mathrm{do}(X_{J\cup C}))$.*

3. *We say that $P(X_B|\,\mathrm{do}(X_{J\cup C}))$ is **trackable up to oracle choices** from $P(X_V|\,\mathrm{do}(X_J))$ (and G) if there exists a conditional Markov kernel at each conditioning step ("chosen by an oracle that knows M") such that following these operations leads to $P(X_B|\,\mathrm{do}(X_{J\cup C}))$.*

4. *Again, we sometimes further restrict the class of CBNs to define/achieve trackability (up to oracle choices), e.g. by adding "for linear Gaussian CBNs" or "for discrete CBNs with strictly positive mass functions", etc., and then only allow M to come from such classes.*

Example 5.3.3. *To illustrate how such a series of operations could look like consider the CADMG G from Figure 12 with $V = \{v_1, v_2, v_3\}$. We assume that the observational distribution $P(X_1, X_2, X_3)$ is given. A list of operations could look like:*

1. *Condition $P(X_1, X_2, X_3)$ on (X_1, X_2) and get a version $P(X_3|X_1, X_2)$.*

   Further specification could be (if possible): "take a continuous version" or "take a version that is only dependent on variables X_1" or "take a strictly positive version".

2. *Marginalize out (X_2, X_3) from $P(X_1, X_2, X_3)$ and get $P(X_1)$.*

3. *Take the product of the previous two Markov kernels: $P(X_3|X_1, X_2) \otimes P(X_1)$.*

4. *Marginalize out X_1 from the last Markov kernel and get: $P(X_3|X_1, X_2) \circ P(X_1)$.*

Lemma 5.3.4. *Let $G = (J, V, E, L)$ be a CADMG.*

1. *If $P(X_B|\,\mathrm{do}(X_{J\cup C}))$ is trackable from $P(X_V|\,\mathrm{do}(X_J))$ then it is also identifiable.*

2. *If $P(X_B|\,\mathrm{do}(X_{J\cup C}))$ is trackable up to oracle choices from $P(X_V|\,\mathrm{do}(X_J))$ then it is also trackable (and thus identifiable) for discrete CBNs M with strictly positive mass functions: $p(x_V|\,\mathrm{do}(x_J)) > 0$ for all x_V, x_J.*

Proof. The first point is clear as the sequence of operations always ends in the same result. For discrete CBNs M with strictly positive mass functions conditional Markov kernels are unambiguous, thus a sequence of marginalization, conditioning and products always leads to the same result. Note that marginals, conditionals and products of strictly positive mass functions also are strictly positive mass functions. $\square$

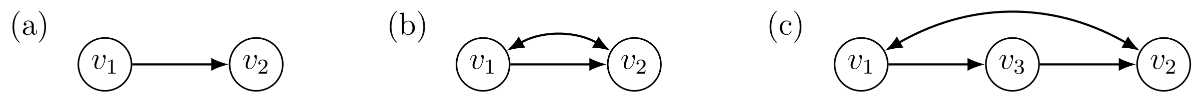


Figure 11: (a) A DAG with two nodes. (b) An ADMG with two nodes. (c) An ADMG with three nodes. The interventional distribution $P(X_2|\,\mathrm{do}(X_1))$ is trackable up to oracle choices from $P(X_1, X_2, X_3)$ in (a) and (c), but not in (b).

Example 5.3.5. *Consider the DAG $G = (V, E)$ from Figure 11 (a) with $V = \{v_1, v_2\}$ and $E = \{v_1 \longrightarrow v_2\}$. Let $\mathcal{X}_1 := \{a, b, c\}$ and $\mathcal{X}_2 := \{0, 1\}$. Define the following Markov kernels $P_1(X_1) := P_1(X_2) := P(X_1)$ via:*

$$P(X_1 = a) := \frac{1}{2}, \qquad P(X_1 = b) := \frac{1}{2}, \qquad P(X_1 = c) := 0,$$

and further:

$$\begin{aligned} P_1(X_2|\,\mathrm{do}(X_1 = a)) &= \mathrm{Bern}(1/4), & P_2(X_2|\,\mathrm{do}(X_1 = a)) &= \mathrm{Bern}(1/4),\\ P_1(X_2|\,\mathrm{do}(X_1 = b)) &= \mathrm{Bern}(3/4), & P_2(X_2|\,\mathrm{do}(X_1 = b)) &= \mathrm{Bern}(3/4),\\ P_1(X_2|\,\mathrm{do}(X_1 = c)) &= \mathrm{Bern}(1/8), & P_2(X_2|\,\mathrm{do}(X_1 = c)) &= \mathrm{Bern}(7/8). \end{aligned}$$

Then we have two CBNs with observable DAG G, the same underlying spaces and the same observable Markov kernel:

$$P(X_1, X_2) := P_1(X_2|\,\mathrm{do}(X_1)) \otimes P_1(X_1) = P_2(X_2|\,\mathrm{do}(X_1)) \otimes P_2(X_1),$$

given by:

$$\begin{aligned} M_1 &:= \left(G, (P_1(X_1), P_1(X_2|\,\mathrm{do}(X_1)))\right),\\ M_2 &:= \left(G, (P_2(X_1), P_2(X_2|\,\mathrm{do}(X_1)))\right). \end{aligned}$$

Furthermore, consider the Markov kernel $P(X_2|X_1)$ given by:

$$\begin{aligned} P(X_2|X_1 = a) &= \mathrm{Bern}(1/4),\\ P(X_2|X_1 = b) &= \mathrm{Bern}(3/4),\\ P(X_2|X_1 = c) &= \mathrm{Bern}(5/8). \end{aligned}$$

We thus have three different versions of the conditional Markov kernels of $P(X_1, X_2)$:

$$P(X_2|X_1) \neq P_1(X_2|\,\mathrm{do}(X_1)) \neq P_2(X_2|\,\mathrm{do}(X_1)) \neq P(X_2|X_1).$$

This shows that the interventional distribution $P(X_2|\,\mathrm{do}(X_1))$ is not identifiable (and thus not trackable) from $P(X_1, X_2)$ and G. However, it is trackable up to oracle choices from $P(X_1, X_2)$ and G. It would thus be trackable (and identifiable) for discrete CBNs with strictly positive mass functions, e.g. here if we also put positive mass on $P(X_1 = c) > 0$.

Notation 5.3.6. *Let $G = (J, V, E, L)$ be a CADMG, $<$ a topological order of G and let $v \in C \subseteq V$. We then put:*

$$\begin{aligned}
\mathrm{Anc}^{[C]}(v) &:= \mathrm{Anc}^{G_{\mathrm{do}(C^\complement)}}(v) \cap C, && \\
\mathrm{Pred}^{[C]}_<(v) &:= \mathrm{Pred}^{G_{\mathrm{do}(C^\complement)}}_<(v) \cap C && = \mathrm{Pred}^{G}_<(v) \cap C, \\
\mathrm{Dist}^{[C]}(v) &:= \mathrm{Dist}^{G_{\mathrm{do}(C^\complement)}}(v) \cap C && = \mathrm{Dist}^{G_{\mathrm{do}(C^\complement)}}(v), \\
\mathrm{Dist}^{[C]}_<(v) &:= \mathrm{Dist}^{[C]}(v) \cap \mathrm{Pred}^{[C]}_<(v), && \\
\mathcal{D}[C] &:= \left\{ \mathrm{Dist}^{[C]}(v) \,\middle|\, v \in C \right\}. &&
\end{aligned}$$

Similarly, if we use the subscript $\leq$ we then also include v. Also note that the dependence on G in the above constructions is implicit.

Notation 5.3.7 (The key interventional Markov kernels)**.** *Let M be an L-CBN with with observable CADMG $G = (J, V, E, L)$ and $C \subseteq V$ any subset. We will abbreviate:*

$$\mathcal{Q}[C] := P(X_C|\,\mathrm{do}(X_{J\cup V\setminus C})) = P(X_C|\,\mathrm{do}(\cancel{X_{(J\cup V)\setminus(\mathrm{Pa}^G(C)\cup C)}}, X_{\mathrm{Pa}^G(C)\setminus C})),$$

where the latter identification comes from Lemma 5.3.8 (using the global Markov property), which is "surely" determined, and not just up to some null-set.

Note that we have the corner cases:

$$\mathcal{Q}[V] = P(X_V|\,\mathrm{do}(X_J)), \qquad \mathcal{Q}[\emptyset] = \delta_*.$$

Lemma 5.3.8. *Let M be an L-CBN with with observable CADMG $G = (J, V, E, L)$ and $C \subseteq V$ any subset. Then we have the identification:*

$$P(X_C|\,\mathrm{do}(X_{J\cup V\setminus C})) = P(X_C|\,\mathrm{do}(\cancel{X_{(J\cup V)\setminus(\mathrm{Pa}^G(C)\cup C)}}, X_{\mathrm{Pa}^G(C)\setminus C})).$$

Proof. This follows from the global Markov property with:

$$C \underset{G_{\mathrm{do}(I_{V\setminus(C\cup \mathrm{Pa}^G(C))}, \mathrm{Pa}^G(C)\setminus C)}}{\overset{d}{\perp}} I_{V\setminus(C\cup\mathrm{Pa}^G(C))} \,|\, \mathrm{Pa}^G(C) \setminus C.$$

To elaborate the latter, let $P := \mathrm{Pa}^G(C) \setminus C$ and $W := V \setminus (C \cup \mathrm{Pa}^G(C))$ and:

$$V' := V \setminus P = C \,\dot\cup\, W, \qquad J' := (J \setminus P) \,\dot\cup\, P, \qquad G' := G_{\mathrm{do}(I_W, P)}.$$

Now consider a walk from a node $c \in C$ to a node $j \in J' \,\dot\cup\, I_W$ in G':

$$\pi : \qquad c \mathrel{*\!\!-\!\!*} \cdots \mathrel{*\!\!-\!\!*} j.$$

If $j \in P$ then the walk is blocked by P at the endnode $j \in P$. So lets assume the case $j \notin P$. Then the walk is of the form:

$$\pi : \qquad c \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast\, w \leftarrow j,$$

with a $w \in W$. So we can write it further as:

$$\pi : \qquad c = c_0 \ast\!\!-\!\!\ast\, c_1 \ast\!\!-\!\!\ast \cdots c_k \ast\!\!-\!\!\ast\, v \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast\, w \leftarrow j,$$

with $c_0, \dots, c_k \in C$ for some $k \geq 0$, and $v \notin C$, the first occuring node not in C (on π from the left). Note that $v = w$ is possible. If the edge $c_k \ast\!\!-\!\!\ast\, v$ is of the form $c_k \leftarrow v$ then $v \in P$ and the walk is blocked by P at the non-collider v. So we can assume the case where the edge is of the form $c_k \ast\!\!\rightarrow v$. This means that on the subwalk $c_k \ast\!\!\rightarrow v \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast\, w \leftarrow j$ we must have at least one collider. This collider is then blocked by P as no collider can be an ancestor of a node in P inside G', because P consists only of input nodes of G'.

This shows the claim:

$$C \underset{G'}{\overset{d}{\perp}} I_W \mid P.$$

The rest then follows from the global Markov property. $\square$

5.3.2. The Interventional Ordered Local Markov Property

One of the ingredient for the ID-algorithm is the ability to track (up to oracle choices) the interventional Markov kernel $\mathcal{Q}[D]$ for districts D of G from $\mathcal{Q}[V]$. The key ingredient to achieve this is the *interventional ordered local Markov property*, which provides us with certain well-behaved Markov kernels that appear in factorizations of both $\mathcal{Q}[V]$ and $\mathcal{Q}[D]$.

Definition 5.3.9 (The preceding Markov blanket of a node)**.** *Let $G = (J, V, E, L)$ be a CADMG and $<$ a topological order. For $v \in V$ we make the following abbreviations:*

$$\begin{aligned}
G_{\leq}(v) &:= \mathrm{Pred}^G_{\leq}(v), \\
\mathrm{Di}^G_{\leq}(v) &:= \mathrm{Dist}^{G_{\leq}(v)}(v), \\
\mathrm{Di}^G_{<}(v) &:= \mathrm{Di}^G_{\leq}(v) \setminus \{v\}, \\
\mathrm{PaD}^G_{<}(v) &:= \mathrm{Pa}^G(\mathrm{Di}^G_{\leq}(v)) \setminus \mathrm{Di}^G_{\leq}(v), \\
\mathrm{Mb}^G_{<}(v) &:= \mathrm{PaD}^G_{<}(v) \,\dot{\cup}\, \mathrm{Di}^G_{<}(v) \\
&= \mathrm{Pa}^G(\mathrm{Dist}^{G_{\leq}(v)}(v)) \cup \mathrm{Dist}^{G_{\leq}(v)}(v) \setminus \{v\}.
\end{aligned}$$

We call $\mathrm{Di}^G_{\leq}(v)$ the ***preceding district*** *and $\mathrm{Mb}^G_{<}(v)$ the* ***preceding Markov blanket*** *of v in G w.r.t. $<$. Note that this definition depends on the topological order $<$ and that we have the inclusions:*

$$\mathrm{Di}^G_{<}(v) \subseteq \mathrm{Mb}^G_{<}(v) \subseteq \mathrm{Pred}^G_{<}(v).$$

Proposition 5.3.10 (Interventional ordered local Markov property)**.** *Let M be an L-CBN with with observable CADMG $G = (J, V, E, L)$ and a fixed topological order $<$. Then for every $v \in V$ we have the conditional independence:*

$$X_v \underset{P(X_V \,|\, \mathrm{do}(X_{I_{V \setminus \mathrm{Di}^G_{\leq}(v)}}, X_J))}{\perp\!\!\!\perp} X_{\mathrm{Pred}^G_{<}(v)} \,|\, X_{\mathrm{Mb}^G_{<}(v)}.$$

In particular, there exists a Markov kernel, denoted by:

$$Q(X_v | X_{\mathrm{Mb}^G_{<}(v)}) \qquad \text{or} \qquad Q(X_v | X_{\mathrm{Di}^G_{<}(v)}, \mathrm{do}(X_{\mathrm{PaD}^G_{<}(v)})),$$

that simultaneously is a version of:

$$P(X_v | X_{\mathrm{Pred}^{[V]}_{<}(v)}, \mathrm{do}(X_J)) \qquad \text{and} \qquad P(X_v | X_{\mathrm{Pred}^{[D]}_{<}(v)}, \mathrm{do}(X_{J \cup V \setminus D})),$$

for every subset $D \subseteq V$ with $\mathrm{Di}^G_{\leq}(v) \subseteq D$, e.g. $D = \mathrm{Dist}^G(v)$.

Proof. This follows from the global Markov property, Theorem 4.2.1, together with the d-separation statement:

$$\{v\} \underset{G_{\mathrm{do}(I_{V \setminus \mathrm{Di}^G_{\leq}(v)})}}{\overset{d}{\perp}} \mathrm{Pred}^G_{<}(v) \,|\, \mathrm{Mb}^G_{<}(v).$$

See Lemma 5.3.12 and Lemma 5.3.13. □

Proofs - The Interventional Ordered Local Markov Property

For the next two Lemmata we introduce some shorter notations:

Notation 5.3.11. *Let $G = (J, V, E, L)$ be a CADMG, $<$ a topological order for G and $v \in V$ a fixed node. For a subset $W \subseteq J \cup V$ we abbreviate:*

$$W_< := \{w \in W \,|\, w < v\}, \qquad W_\leq := \{w \in W \,|\, w < v \,\vee\, w = v\},$$

and $W_>$ and $W_\geq$ accordingly.

The next Lemma is the graphical center piece that makes the ID-algorithm possible. It could be called the graphical version of an "interventional ordered local Markov property", whose distributional counterpart is stated in the Lemma after.

Lemma 5.3.12. *Let $G = (J, V, E, L)$ be a CADMG, $<$ a topological order for G and $v \in V$ a fixed node. Let $D \subseteq V$ be a subset such that $v \in D$ and $\mathrm{Dist}^{G_\leq}(D_\leq) \subseteq D$, where $G_\leq := \mathrm{Pred}^G_{\leq}(v)$ is the ancestral subgraph of predecessors of v in G, e.g. $D = \mathrm{Dist}^G(v)$ or $D = \mathrm{Dist}^{G_\leq}(v)$. Then we have the d-separation statement:*

$$\{v\} \underset{G_{\mathrm{do}(I_{V \setminus D_\leq})}}{\overset{d}{\perp}} V_< \,|\, \mathrm{Pa}^G(D_\leq) \cup D_<.$$

Proof. We abbreviate:

$$\tilde{G} := G_{\mathrm{do}(I_{V\setminus D_{\leq}})}, \qquad F := \mathrm{Pa}^G(D_{\leq}) \cup D_{<}.$$

Assume the contrary to the claim and let π be a shortest path from v to a node $w \in J \dot{\cup} I_{V\setminus D} \dot{\cup} V_{<}$ in the graph $\tilde{G}$. It is clear that $w \neq v$.

If $w \in F$ then π is blocked by F at the endnode w. So we can assume that $w \notin F$, in particular, $w \notin D_{\leq}$. So there exist $v_0, \dots, v_k \in D_{\leq}$ for some $k \geq 0$ and $\tilde{w} \notin D_{\leq}$ such that π is of the form:

$$\pi : \qquad v = v_0 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast v_k \ast\!\!-\!\!\ast \tilde{w} \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast w,$$

where $\tilde{w}$ is the first node from the left that is not in $D_{\leq}$, which exists since $w \notin D_{\leq}$. Since $v_k \in D_{\leq}$ it is clear that $\tilde{w} \notin I_{V\setminus D_{\leq}}$. So $\tilde{w} \in J \cup V \setminus D_{\leq}$.

If the edge $v_k \ast\!\!-\!\!\ast \tilde{w}$ is of the form $v_k \leftarrow \tilde{w}$ then $\tilde{w} \in \mathrm{Pa}^G(D_{\leq})$. So in this case π is blocked at the non-collider $\tilde{w}$ by F.

So we can assume the case $v_k \ast\!\!\rightarrow \tilde{w}$. This implies that $\tilde{w}$ cannot lie in the set of input nodes of $\tilde{G}$, which implies that $\tilde{w} \notin J \cup I_{V\setminus D_{\leq}}$. With this we then get that $\tilde{w} \in V \setminus D_{\leq}$.

Assume the case that $\tilde{w} \in V_{>}$ and π is of the form:

$$\pi : \qquad v = v_0 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast v_k \ast\!\!\rightarrow \tilde{w} = \tilde{w}_0 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast \tilde{w}_m = w,$$

where the subwalk $\tilde{w}_0 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast \tilde{w}_m$ has no colliders. Because of the edge $v_k \ast\!\!\rightarrow \tilde{w}$ this subwalk necessarily is directed to the right and we get:

$$\pi : \qquad v = v_0 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast v_k \ast\!\!\rightarrow \tilde{w} = \tilde{w}_0 \rightarrow \cdots \rightarrow \tilde{w}_m = w.$$

Then the endnode w is not an input node, $w \notin J \cup I_{V\setminus D_{\leq}}$, and $v < \tilde{w} \leq \tilde{w}_m = w$, thus $w \in V_{>}$. But this is a contradiction to $w \in J \dot{\cup} I_{V\setminus D} \dot{\cup} V_{<}$. So this case cannot occur.

Now assume the case that $\tilde{w} \in V_{>}$ and π is of the form:

$$\pi : \qquad v = v_0 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast v_k \ast\!\!\rightarrow \tilde{w} = \tilde{w}_0 \rightarrow \cdots \rightarrow \tilde{w}_m \leftarrow\!\!\ast \cdots \ast\!\!-\!\!\ast w,$$

for some $m \geq 0$, with a directed subwalk $\tilde{w}_0 \rightarrow \cdots \rightarrow \tilde{w}_m$, where $\tilde{w}_m$ is the first node after $\tilde{w}$ where a collider occurs, which could be $\tilde{w}$ itself. Again we have $v < \tilde{w} \leq \tilde{w}_m$ and thus $\tilde{w}_m \in V_{>}$. This implies that $\tilde{w}_m \notin \mathrm{Anc}^G(F)$, since $\mathrm{Anc}^G(F) \subseteq J \cup V_{<}$. So π is blocked by F at the collider $\tilde{w}_m$.

So we are left with the cases $v_k \ast\!\!\rightarrow \tilde{w}$ and $\tilde{w} \in V_{<} \setminus D_{\leq}$.

Now consider the case of a directed edge $v_k \rightarrow \tilde{w}$ and $\tilde{w} \in V_{<} \setminus D_{\leq}$. If $v_k \neq v$ then π is blocked at the non-collider $v_k \in D_{<}$ by F. So we can assume that $v_k = v$. This implies $v < \tilde{w}$ and thus $\tilde{w} \in V_{>}$, which contradicts $\tilde{w} \in V_{<}$. So this cannot occur.

Now consider the case of a bidirected edge $v_k \leftrightarrow \tilde{w}$ and $\tilde{w} \in V_{<} \setminus D_{\leq}$. Then both nodes $v_k, \tilde{w} \in G_{\leq}$ and $\tilde{w} \in \mathrm{Dist}^{G_{\leq}}(v_k)$. By the assumption of this Lemma we have:

$$\tilde{w} \in \mathrm{Dist}^{G_{\leq}}(v_k) \subseteq \mathrm{Dist}^{G_{\leq}}(D_{\leq}) \subseteq D \cap G_{\leq} = D_{\leq}.$$

So $\tilde{w} \in D_{\leq}$, which contradicts $\tilde{w} \notin D_{\leq}$. So this case cannot neither occur.

So we have shown that in all cases that can occur the path π in $\tilde{G}$ is blocked by F. This shows the claim. $\square$

The last Lemma allows us to use the global Markov property for the existence of special Markov kernels that are of importance for the ID-algorithm:

Lemma 5.3.13 (Interventional ordered local Markov property)**.** *Let M be an L-CBN with with observable CADMG $G = (J, V, E, L)$ and a fixed topological order $<$ and fixed $v \in V$. Let $G_{\leq} := \mathrm{Pred}^G_{\leq}(v)$ be the ancestral sub-CADMG of predecessors of v in G and let:*

$$D_{\leq} := \mathrm{Dist}^{G_{\leq}}(v), \qquad F := \mathrm{Pa}^G(D_{\leq}) \cup D_{\leq} \setminus \{v\}\,.$$

Then we have the conditional independence:

$$X_v \underset{P(X_{V_{\leq}}\,|\,\mathrm{do}(X_{I_{V\setminus D_{\leq}}}, X_J))}{\perp\!\!\!\perp} X_{V_<} \,|\, X_F.$$

In particular, there exists a Markov kernel:

$$Q(X_v|X_F)$$

that simultaneously is a version of:

$$P(X_v|X_H, \mathrm{do}(X_{J\cup V\setminus S})),$$

for every subsets $H, S \subseteq V$ such that $D_{\leq} \subseteq S \subseteq V$ and $F \cap S_< \subseteq H \subseteq V_<$. Note that this includes the corner cases:

1. $P(X_v|X_{V_<}, \mathrm{do}(X_J))$,
2. $P(X_v|X_{F\cap V_<}, \mathrm{do}(X_J))$,
3. $P(X_v|X_{S_<}, \mathrm{do}(X_{J\cup V\setminus S}))$,
4. $P(X_v|X_{F\cap S_<}, \mathrm{do}(X_{J\cup V\setminus S}))$,
5. $P(X_v|X_{D_<}, \mathrm{do}(X_{J\cup V\setminus D_{\leq}}))$.

Proof. By Lemma 5.3.12 we have the d-separation:

$$\{v\} \underset{G_{\mathrm{do}(I_{V\setminus D_{\leq}})}}{\overset{d}{\perp}} V_< \,|\, F.$$

By the global Markov property, Theorem 4.2.1, we get:

$$X_v \underset{P(X_{V_{\leq}}\,|\,\mathrm{do}(X_{I_{V\setminus D_{\leq}}}, X_J))}{\perp\!\!\!\perp} X_{V_<} \,|\, X_F.$$

So there exists a Markov kernel $Q(X_v|X_F)$ such that:

$$P(X_{V_{\leq}}|\,\mathrm{do}(X_{I_{V\setminus D_{\leq}}}, X_J)) = Q(X_v|X_F) \otimes P(X_{V_<}|\,\mathrm{do}(X_{I_{V\setminus D_{\leq}}}, X_J)). \qquad (\#)$$

For a subset $S \subseteq V$ with $D_\leq \subseteq S$ we have:

$$V \setminus D_\leq = (V \setminus S) \,\dot\cup\, (S \setminus D_\leq).$$

By putting $X_{I_{S\setminus D_\leq}} = \star$ and $X_{I_{V\setminus S}} = x_{V\setminus S}$ we get:

$$P(X_{V_\leq} |\, \mathrm{do}(X_{V\setminus S}, X_J)) = Q(X_v|X_F) \otimes P(X_{V_<} |\, \mathrm{do}(X_{V\setminus S}, X_J)).$$

Now consider another subset $H \subseteq V_<$ with $F \cap S_< \subseteq H$. Then marginalizing out $X_{V_\leq \setminus (\{v\} \cup H)}$ gives us:

$$P(X_v, X_H |\, \mathrm{do}(X_{J\cup V\setminus S})) = Q(X_v|X_F) \otimes P(X_H |\, \mathrm{do}(X_{J\cup V\setminus S})).$$

This shows that $Q(X_v|X_F)$, simultaneously, is a version of:

$$P(X_v|X_H, \mathrm{do}(X_{J\cup V\setminus S})),$$

for every subsets $H, S \subseteq V$ such that $D_\leq \subseteq S \subseteq V$ and $F \cap S_< \subseteq H \subseteq V_<$. This shows the claim. $\square$

5.3.3. Ancestral Sets and Districts

Lemma 5.3.14 (Ancestral subsets are trackable)**.** *Let M be an L-CBN with with observable CADMG $G = (J, V, E, L)$ and $A \subseteq V$ be a subset such that $A = \mathrm{Anc}^{[V]}(A)$. Then we have the equality between the interventional distribution $\mathcal{Q}[A]$ and the A-marginal of $\mathcal{Q}[V]$:*

$$\mathcal{Q}[A] = P(X_A |\, \mathrm{do}(X_{J\cup V\setminus A})) = P(X_A |\, \mathrm{do}(X_J)).$$

Proof. By Lemma 5.3.8 we only have to show the right identity:

$$\mathcal{Q}[A] = P(X_A |\, \mathrm{do}(X_{J\cup V\setminus A})) \stackrel{!}{=} P(X_A |\, \mathrm{do}(X_J)).$$

The latter follows again from the global Markov property and the d-separation:

$$A \underset{G_{\mathrm{do}(I_{V\setminus A})}}{\overset{d}{\perp}} I_{V\setminus A} \,|\, J.$$

This d-separation holds true since every walk from a node $v \in A$ to a node $j \in J \,\dot\cup\, I_{V\setminus A}$ is either blocked by J as the endnode $j \in J$ or is of the form:

$$\pi: \qquad v \ast\!\!-\!\!\ast\, a_1 \ast\!\!-\!\!\ast \cdots a_k \ast\!\!-\!\!\ast\, w' \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast\, w \leftarrow j,$$

with a $w \in V \setminus A$, $j \in I_{V\setminus A}$, $a_1, \dots, a_k \in A$ for some $k \geq 0$ and $w' \notin A$, the first node not in A on π from the left ($w' = w$ possible). In case the edge $a_k \ast\!\!-\!\!\ast\, w'$ is of the form $a_k \leftarrow w'$ we have:

$$w' \in \mathrm{Anc}^G(A) \setminus A = \mathrm{Anc}^G(A) \setminus \mathrm{Anc}^{[V]}(A) \subseteq J.$$

So in this case the walk is blocked at the non-collider w' by J. So we can consider the case where the edge is of the form $a_k \ast\!\!\rightarrow w'$. Then the subwalk $a_k \ast\!\!\rightarrow w' \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast\, w \leftarrow j$ must contain a collider. This collider can not be an ancestor of J, as J are the input nodes. So the walk is blocked by J in all cases. $\square$

Remark 5.3.15 (Districts are trackable up to oracle choices). *Let M be an L-CBN with observable CADMG $G = (J, V, E, L)$.*

1. *Since the Markov kernel $Q(X_v|X_{\mathrm{Mb}^{\mathcal{G}}_{<}(v)})$ coming from the interventional ordered local Markov property, see Proposition 5.3.10, is a version of both:*

$$P(X_v|X_{\mathrm{Pred}^{[V]}_{<}(v)}, \mathrm{do}(X_J)) \qquad \textit{and} \qquad P(X_v|X_{\mathrm{Pred}^{[D]}_{<}(v)}, \mathrm{do}(X_{J\cup V\setminus D})),$$

   for $D = \mathrm{Dist}^G(v)$, which are marginal conditionals of the interventional distributions $\mathcal{Q}[V]$ and $\mathcal{Q}[D]$, resp., the $Q(X_v|X_{\mathrm{Mb}^{\mathcal{G}}_{<}(v)})$'s are trackable from either quantity up to oracle choices.

2. *We get the following factorization by the chain rule for every $D \in \mathcal{D}[V]$:*

$$\mathcal{Q}[V] = \bigotimes_{v\in V}^{>} P(X_v|X_{\mathrm{Pred}^{[V]}_{<}(v)}, \mathrm{do}(X_J)) \qquad = \bigotimes_{v\in V}^{>} Q(X_v|X_{\mathrm{Mb}^{\mathcal{G}}_{<}(v)}),$$
$$\mathcal{Q}[D] = \bigotimes_{v\in D}^{>} P(X_v|X_{\mathrm{Pred}^{[D]}_{<}(v)}, \mathrm{do}(X_{J\cup V\setminus D})) \qquad = \bigotimes_{v\in D}^{>} Q(X_v|X_{\mathrm{Mb}^{\mathcal{G}}_{<}(v)}).$$

3. *In particular, $\mathcal{Q}[D]$ is trackable from $\mathcal{Q}[V]$ up to oracle choices by first determining $Q(X_v|X_{\mathrm{Mb}^{\mathcal{G}}_{<}(v)})$ for $v \in D$ via marginalization and conditioning and then taking the product of Markov kernels in reverse order of $<$.*

4. *The above seems to give us something like a factorization:*

$$\mathcal{Q}[V] = \bigotimes_{v\in V}^{>} Q(X_v|X_{\mathrm{Mb}^{\mathcal{G}}_{<}(v)}) \qquad = \left[\bigotimes_{D\in\mathcal{D}[V]}^{>} \bigotimes_{v\in D}^{>}\right] Q(X_v|X_{\mathrm{Mb}^{\mathcal{G}}_{<}(v)})$$
$$= \left[\bigotimes_{D\in\mathcal{D}[V]}^{>}\right] \left(\bigotimes_{v\in D}^{>} Q(X_v|X_{\mathrm{Mb}^{\mathcal{G}}_{<}(v)})\right) \qquad = \left[\bigotimes_{D\in\mathcal{D}[V]}^{>}\right] \mathcal{Q}[D],$$

   where the products in brackets are not well-defined in the naive way, as the districts of a CADMG don't need to be topologically ordered. Nonetheless, if we are given a fixed topological order $<$ and the interventional Markov kernels $\mathcal{Q}[D]$ for every $D \in \mathcal{D}[V]$ then we can first track $Q(X_v|X_{\mathrm{Mb}^{\mathcal{G}}_{<}(v)})$ from $\mathcal{Q}[D]$ up to oracle choices for every $v \in D$ and every $D \in \mathcal{D}[V]$ and then take the product on the top left.

Definition 5.3.16. *Let M be an L-CBN with observable CADMG $G = (J, V, E, L)$ and a fixed topological order $<$. Let $\mathcal{D} \subseteq \mathcal{D}[V]$ be a set of districts of G. Then we put:*

$$\left[\bigotimes_{D\in\mathcal{D}}^{>}\right] \mathcal{Q}[D] := \bigotimes_{v\in\bigcup_{D\in\mathcal{D}} D}^{>} Q(X_v|X_{\mathrm{Mb}^{\mathcal{G}}_{<}(v)}),$$

where the product on the right is taken in reverse topological order. Note that for $G' := G_{\mathrm{do}(D^c)}$ and $v \in D$ we have $\mathrm{Mb}^{G'}_{<}(v) = \mathrm{Mb}^{\mathcal{G}}_{<}(v)$. So the set $\mathrm{Mb}^{\mathcal{G}}_{<}(v)$ can be determined by the subgraph G' and $<$ alone.

5.3.4. The ID-Algorithm

Now we come to the main part of this section, the *ID-algorithm* for the identification of causal effects, or, more precisely, the trackability up to oracle choices of interventional Markov kernels, from the observable Markov kernel. The main references are [Pea09, GP95, Tia02, TP02, Tia04, SP06b, HV06, HV08, RERS23, FM20].

Algorithm 5.3.17 (ID-algorithm)**.** *Let M be an L-CBN with with observable CADMG $G = (J, V, E, L)$ and a fixed topological order $<$. Let $\emptyset \neq B \subseteq V$ and $C \subseteq J \cup V$ be two disjoint subsets of nodes. We want to query if the interventional Markov kernel $P(X_B|\,\mathrm{do}(X_{J\cup C}))$ is trackable up to oracle choices from the observable Markov kernel $P(X_V|\,\mathrm{do}(X_J)) = \mathcal{Q}[V]$ and G.*

1. *Put $B^C := \mathrm{Anc}^{G_{\mathrm{do}(C)}}(B) \setminus (J \cup C) \subseteq V$.*

   Then $P(X_B|\,\mathrm{do}(X_{J\cup C}))$ is the B-marginal of $P(X_{B^C}|\,\mathrm{do}(X_{J\cup C})) = \mathcal{Q}[B^C]$.

   So we are left to determine if we can track $\mathcal{Q}[B^C]$ from $\mathcal{Q}[V]$ up to oracle choices.

2. *Find the districts $\mathcal{D}[B^C] = \{S_1, \dots, S_K\}$ and put $A_{k,0} := V$ for $k = 1, \dots, K$.*

   Note that $\mathcal{Q}[A_{k,0}] = \mathcal{Q}[V]$ is trivially tracked from $\mathcal{Q}[V]$.

3. *For each $k = 1, \dots, K$ repeat the following steps recursively for $\ell \in \mathbb{N}$:*
   a) *Take the district in $A_{k,\ell}$:*

   $$D_{k,\ell} := \mathrm{Dist}^{[A_{k,\ell}]}(S_k)$$

   We can track $\mathcal{Q}[D_{k,\ell}]$ from $\mathcal{Q}[A_{k,\ell}]$ up to oracle choices by Remark 5.3.15.

   b) *Take the ancestral closure in $D_{k,\ell}$:*

   $$A_{k,\ell+1} := \mathrm{Anc}^{[D_{k,\ell}]}(S_k).$$

   We can track $\mathcal{Q}[A_{k,\ell+1}]$ from $\mathcal{Q}[D_{k,\ell}]$ via marginalization by Lemma 5.3.14.

   c) *If $D_{k,\ell} = A_{k,\ell}$ or $A_{k,\ell+1} = D_{k,\ell}$ then stop for this k and put:*

   $$\breve{S}_k := D_{k,\ell}.$$

   Otherwise, repeat with: $\ell \leftarrow \ell + 1$.

4. *When the algorithm has stopped then for every $k = 1, \dots, K$ we have:*

   $$\breve{S}_k = \mathrm{Anc}^{[\breve{S}_k]}(S_k) = \mathrm{Dist}^{[\breve{S}_k]}(S_k) \supseteq S_k.$$

   Furthermore, we have tracked all $\mathcal{Q}[\breve{S}_k]$'s recursively from $\mathcal{Q}[V]$ up to oracle choices.

5. *If there is any $k = 1, \dots, K$ with $\breve{S}_k \neq S_k$ then the ID-algorithm outputs: FAIL.*

6. *Otherwise, we have for all $k = 1, \dots, K$ that $\mathcal{Q}[S_k] = \mathcal{Q}[\breve{S}_k]$ and we can track $\mathcal{Q}[B^C]$ from $\mathcal{Q}[V]$ up to oracle choices via:*

$$\mathcal{Q}[B^C] = \left[\bigotimes_{S_k \in \mathcal{D}[B^C]}^{>} \right] \mathcal{Q}[S_k],$$

and $P(X_B|\,\mathrm{do}(X_{J\cup C}))$ as the B-marginal thereof.

Corollary 5.3.18 (Soundness up to oracle choices). *The ID-algorithm 5.3.17 is sound up to oracle choices. This means that if it does not produce FAIL for input $B, C \subseteq G$ then $P(X_B|\,\mathrm{do}(X_{J\cup C}))$ is trackable up to oracle choices from $P(X_V|\,\mathrm{do}(X_J))$ and G.*

Proof. This is clear as each step in the ID-algorithm is trackable up to oracle choices. Note that the operations at each step can be formulated by knowing G, B and C alone without knowing M in advance. □

Theorem 5.3.19 (Soundness up to null-sets). *Let $G = (J, V, E, L)$ be a CADMG with a fixed topological order $<$. Consider the class of L-CBNs M with observable CADMG G such that the following holds:*

1. *Each measurable space $\mathcal{X}_v$ comes equipped with a fixed measure μ_v for $v \in V$,*

2. *for every subset $D \subseteq V$ the interventional Markov kernel $\mathcal{Q}[D] = P(X_D|\,\mathrm{do}(X_{J\cup V\setminus D}))$ is **absolute continuous** w.r.t. the product measure $\mu_D := \bigotimes_{v\in D} \mu_v$ and vice versa:*

$$\mu_D \ll \mathcal{Q}[D] \ll \mu_D.$$

If the ID-algorithm does not produce FAIL for input $B, C \subseteq G$, then $P(X_B|\,\mathrm{do}(X_{J\cup C}))$ is "almost-surely" trackable from $P(X_V|\,\mathrm{do}(X_J))$ and G for such CBNs M, i.e. the Markov kernel that was output by the ID-algorithm equals $P(X_B|\,\mathrm{do}(X_{J\cup C}))$ up to a $\mu_{V\setminus B^C}$-null set in $\mathcal{X}_{J\cup V\setminus B^C}$ (see Remark 5.3.20 below).

Proof. See Theorem 5.3.31. □

Remark 5.3.20.

1. *For the almost-sure soundness in Theorem 5.3.19 to hold one implicitely needs/is allowed to make slight relaxations to the ID-algorithm 5.3.17:*

   Instead of insisting on taking the conditionals $Q(X_v|X_{\mathrm{Mb}^{G'}_{<}(v)})$ of $\mathcal{Q}[D]$ for $v \in D \subseteq V$ one takes any version of that conditional of $\mathcal{Q}[D] \otimes \mu_{V\setminus D}$ that is only dependent on predecessors of v, which will always be possible as $Q(X_v|X_{\mathrm{Mb}^{G'}_{<}(v)})$ is an existing such version.

2. *The conditions of Theorem 5.3.19 are satisfied for a L-CBNs M with CDAG $G^+ = (J, U \dot\cup V, E)$ if every Markov kernel $P_v(X_v|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)}))$ has a strictly positive Doob-Radon-Nikodym derivative/density w.r.t. the measure μ_v for all $v \in V$, see Lemma 5.3.33.*

Theorem 5.3.21 (From almost-sure to sure soundness)**.** *In addition to the conditions in Theorem 5.3.19 assume:*

1. *$\mathcal{X}_v$ is a **Polish space** for every $v \in J \cup V$,*
2. *μ_V is **strictly positive** (on non-empty open subsets of $\mathcal{X}_V$),*
3. *the queried interventional Markov kernel is **continuous** as a map:*

$$P(X_B|\operatorname{do}(X_{J\cup C})) : \mathcal{X}_{J\cup C} \to \mathcal{P}(\mathcal{X}_B).$$

Then the output $\hat{P}(X_B|X_{J\cup V\setminus B^C})$ of the ID-algorithm (in the not-FAIL case) can be changed on a $\mu_{V\setminus B^C}$-null set in $\mathcal{X}_{J\cup V\setminus B^C}$ such that it becomes continuous as a map:

$$\hat{P}(X_B|X_{J\cup V\setminus B^C}) : \mathcal{X}_{J\cup V\setminus B^C} \to \mathcal{P}(\mathcal{X}_B).$$

Every such continuous version of $\hat{P}(X_B|X_{J\cup V\setminus B^C})$ is then necessarily identical to the interventional Markov kernel $P(X_B|\operatorname{do}(X_{J\cup C}))$. These conditions thus allow us to recover from the ambiguity resulting from the null-sets.

Proof. The existence of such a null-set is clear, because $\hat{P}(X_B|X_{J\cup V\setminus B^C})$ is $\mu_{V\setminus B^C}$-almost-surely equal to $P(X_B|\operatorname{do}(X_{J\cup C}))$, and the latter was assumed to be continuous. The uniqueness follows from Lemma 2.4.23. □

Remark 5.3.22. *The statement of Theorem 5.3.21 can be further relaxed by asking for Polish spaces $\mathcal{X}_v$ only for $v \in V$, strict positivity only for $\mu_{V\setminus B^C}$ and only for the continuity of the maps:*

$$\mathcal{X}_{C\setminus J} \to \mathcal{P}(\mathcal{X}_B), \quad x_{C\setminus J} \mapsto P(X_B|\operatorname{do}(X_{J\cup C} = (x_J, x_{C\setminus J})),$$

for every $x_J \in \mathcal{X}_J$ separately, by then applying the criterion from Theorem 5.3.21 for each partial input $x_J \in \mathcal{X}_J$ separately.

Example 5.3.23. *All stated assumptions of Theorems 5.3.19 and 5.3.21 are satisfied if every Markov kernel of the L-CBN M is linear Gaussian:*

$$P_v(X_v \in dx_v|\operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)} = x_{\mathrm{Pa}^{G^+}(v)})) = \mathcal{N}(dx_v|\Gamma_v \cdot x_{\mathrm{Pa}^{G^+}(v)} + \gamma_v, \Sigma_v),$$

with transition matrix Γ_v, translation vector γ_v and positive definite covariance matrix $\Sigma_v \succ 0$ and Lebesgue measures μ_v, $v \in U \cup V$.

Theorem 5.3.24 (Completeness, see [HV08])**.** *The ID-algorithm is complete.*

More precisely, if the ID-algorithm outputs FAIL for subsets $B, C \subseteq G = (J, V, E, L)$ then there exist two L-CBNs M_1 and M_2 with the same observable CADMG G, the same and discrete underlying spaces $\mathcal{X}_v$ for $v \in J\cup V$, and the same observable Markov kernels $P_1(X_V|\operatorname{do}(X_J)) = P_2(X_V|\operatorname{do}(X_J))$ that have strictly positive mass functions such that:

$$P_1(X_B|\operatorname{do}(X_{J\cup C})) \neq P_2(X_B|\operatorname{do}(X_{J\cup C})).$$

In particular, in case of FAIL, $P(X_B|\operatorname{do}(X_{J\cup C}))$ is not identifiable from $P(X_V|\operatorname{do}(X_J))$ and G.

Remark 5.3.25 (Identification of conditional causal effects, see [Tia04]). *If we want to know if the conditional interventional Markov kernel $P(X_A|X_B, \mathrm{do}(X_{J\cup C}))$ is trackable up to oracle choices from $P(X_V|\,\mathrm{do}(X_J))$ and G then we can run the ID-algorithm for $A \cup B$ and C. If it does not output FAIL then $P(X_{A\cup B}|\,\mathrm{do}(X_{J\cup C}))$ is trackable up to oracle choices from $P(X_V|\,\mathrm{do}(X_J))$ and G and by conditioning on X_B afterwards so will $P(X_A|X_B, \mathrm{do}(X_{J\cup C}))$ be.*

However, note that there is a "conditional" version of the ID-algorithm, see [Tia04], that can check if (and conclude that) $P(X_A|X_B, \mathrm{do}(X_{J\cup C}))$ is trackable up to oracle choices from $P(X_V|\,\mathrm{do}(X_J))$ and G even if the (unconditional) ID-algorithm outputs FAIL for $P(X_{A\cup B}|\,\mathrm{do}(X_{J\cup C}))$.

Examples

Example 5.3.26. *Consider the DAG $G = (V, E)$ from Figure 11 (a) with $V = \{v_1, v_2\}$, $E = \{v_1 \longrightarrow v_2\}$. We want to determine if we can identify $P(X_2|\,\mathrm{do}(X_1))$ from $P(X_1, X_2)$ in case we have a discrete CBN with strictly positive mass function: $p(x_1, x_2, x_3) > 0$. For this let $B := \{v_2\}$ and $C := \{v_1\}$. Note that we have the topological order $v_1 < v_2$. We then follow the steps of the ID-algorithm:*

1. $\mathcal{Q}[V](x_1, x_2) := p(x_1, x_2)$.
2. $B^C = \mathrm{Anc}^{G_{\mathrm{do}(C)}}(B) \setminus C = \{v_2\}$.
3. $\mathcal{D}[B^C] = \{S = \{v_2\}\}$. *So we compute:*
   a) $D_0 = \mathrm{Dist}^{[V]}(S) = \{v_2\}$. *Compute:*
   $$\mathcal{Q}[D_0](x_2|x_1) = q(x_2|x_1) = \frac{\mathcal{Q}[V](x_1, x_2)}{\mathcal{Q}[V](x_1)} = p(x_2|x_1).$$
   b) $A_1 = \mathrm{Anc}^{[D_0]}(S) = \{v_2\} = D_0$, *thus* $\breve{S} = D_0 = \{v_2\} = S$. *Compute:*
   $$\mathcal{Q}[S](x_2|x_1) = \mathcal{Q}[D_0](x_2|x_1) = p(x_2|x_1).$$
4. *Since* $\breve{S} = S = \{v_2\}$ *we can compute:*
   $$\mathcal{Q}[B^C](x_2|x_1) = \mathcal{Q}[S_1](x_2|x_1) = p(x_2|x_1).$$

So we can identify $P(X_2|\,\mathrm{do}(X_1))$ from $P(X_1, X_2)$ as the conditional $P(X_2|X_1)$ via the mass function from above.

Example 5.3.27. *Consider the ADMG $G = (V, E, L)$ from Figure 11 (b) with $V = \{v_1, v_2\}$, $E = \{v_1 \longrightarrow v_2\}$ and $L = \{v_1 \longleftrightarrow v_2\}$. We want to determine if we can identify $P(X_2|\,\mathrm{do}(X_1))$ from $P(X_1, X_2)$ in case we have a discrete CBN with strictly positive mass function: $p(x_1, x_2, x_3) > 0$. For this let $B := \{v_2\}$ and $C := \{v_1\}$. Note that we have the topological order $v_1 < v_2$. We then follow the steps of the ID-algorithm:*

1. $\mathcal{Q}[V](x_1, x_2) := p(x_1, x_2)$.
2. $B^C = \mathrm{Anc}^{G_{\mathrm{do}(C)}}(B) \setminus C = \{v_2\}$.
3. $\mathcal{D}[B^C] = \{S = \{v_2\}\}$. *So we compute:*
   a) $D_0 = \mathrm{Dist}^{[V]}(S) = \{v_1, v_2\}$. *Compute:*

$$\begin{aligned} q(x_1) &= \mathcal{Q}[V](x_1) && = p(x_1), \\ q(x_2|x_1) &= \frac{\mathcal{Q}[V](x_1, x_2)}{\mathcal{Q}[V](x_1)} && = p(x_2|x_1), \\ \mathcal{Q}[D_0](x_1, x_2) &= q(x_2|x_1) \cdot q(x_1) && = p(x_1, x_2). \end{aligned}$$

   b) $A_1 = \mathrm{Anc}^{[D_0]}(S) = \{v_1, v_2\} = D_0$, *thus* $\breve{S} = D_0 = \{v_1, v_2\}$. *Compute:*

$$\mathcal{Q}[\breve{S}](x_1, x_2) = \mathcal{Q}[D_0](x_1, x_2) = p(x_1, x_2).$$

4. *Since* $\breve{S} = \{v_1, v_2\} \neq \{v_2\} = S$ *the ID-algorithms outputs: FAIL.*

So we can not identify $P(X_2|\,\mathrm{do}(X_1))$ *from* $P(X_1, X_2)$ *and* G.

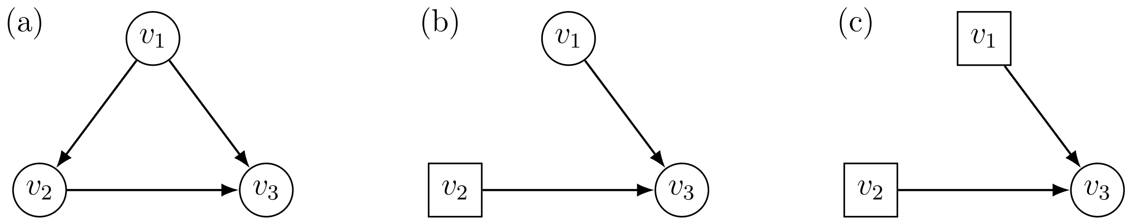


Figure 12: A DAG with three nodes and its intervened graphs.

Example 5.3.28. *Consider the DAG* $G = (V, E, L)$ *from Figure 12 with* $V = \{v_1, v_2, v_3\}$, $E = \{v_1 \longrightarrow v_2, v_1 \longrightarrow v_3, v_2 \longrightarrow v_3\}$. *We want to determine if we can identify* $P(X_3|\,\mathrm{do}(X_2))$ *from* $P(X_1, X_2, X_3)$ *in case we have a discrete CBN with strictly positive mass function:* $p(x_1, x_2, x_3) > 0$. *For this let* $B := \{v_3\}$ *and* $C := \{v_2\}$. *Note that we have the topological order* $v_1 < v_2 < v_3$. *We then follow the steps of the ID-algorithm:*

1. $\mathcal{Q}[V](x_1, x_2, x_3) := p(x_1, x_2, x_3)$.
2. $B^C = \mathrm{Anc}^{G_{\mathrm{do}(C)}}(B) \setminus C = \{v_1, v_3\}$.
3. $\mathcal{D}[B^C] = \{S_1 = \{v_1\}, S_2 = \{v_3\}\}$, *see Figure 12 (b).*
4. *For* $S_1 = \{v_1\}$:
   a) $D_{1,0} = \mathrm{Dist}^{[V]}(S_1) = \{v_1\}$. *Compute:*

$$\mathcal{Q}[D_{1,0}](x_1) = q(x_1) = \mathcal{Q}[V](x_1) = p(x_1).$$

b) $A_{1,1} = \mathrm{Anc}^{[D_{1,0}]}(S_1) = \{v_1\} = D_{1,0}$, *thus* $\breve{S}_1 = D_{1,0}$. *Compute:*

$$\mathcal{Q}[\breve{S}_1](x_1) = \mathcal{Q}[D_{1,0}](x_1) = p(x_1).$$

c) $S_1 = \{v_1\} = \breve{S}_1$. *Compute:*

$$\mathcal{Q}[S_1](x_1) = \mathcal{Q}[\breve{S}_1](x_1) = p(x_1).$$

5. *For* $S_2 = \{v_3\}$*:*

a) $D_{2,0} = \mathrm{Dist}^{[V]}(S_2) = \{v_3\}$. *Compute:*

$$\mathcal{Q}[D_{2,0}](x_3|x_1,x_2) = q(x_3|x_1,x_2) = \frac{\mathcal{Q}[V](x_1,x_2,x_3)}{\mathcal{Q}[V](x_1,x_2)} = p(x_3|x_1,x_2).$$

b) $A_{2,1} = \mathrm{Anc}^{[D_{2,0}]}(S_1) = \{v_3\} = D_{2,0}$. *thus* $\breve{S}_2 = D_{2,0}$. *Compute:*

$$\mathcal{Q}[\breve{S}_2](x_3|x_1,x_2) = \mathcal{Q}[D_{2,0}](x_3|x_1,x_2) = p(x_3|x_1,x_2).$$

c) $S_2 = \{v_3\} = \breve{S}_2$. *Compute:*

$$\mathcal{Q}[S_2](x_3|x_1,x_2) = \mathcal{Q}[\breve{S}_2](x_3|x_1,x_2) = p(x_3|x_1,x_2).$$

6. *Since both* $\breve{S}_1 = S_1 = \{v_1\}$ *and* $\breve{S}_2 = S_2 = \{v_3\}$ *we can compute:*

$$\begin{aligned} \mathcal{Q}[B^C](x_1,x_3|x_2) &= \mathcal{Q}[S_2](x_3|x_1,x_2) \cdot \mathcal{Q}[S_1](x_1) \\ &= p(x_3|x_1,x_2) \cdot p(x_1), \\ p(x_3|\,\mathrm{do}(x_2)) &= \sum_{x_1} \mathcal{Q}[B^C](x_1,x_3|x_2) \\ &= \sum_{x_1} p(x_3|x_1,x_2) \cdot p(x_1). \end{aligned}$$

So we can identify $P(X_3|\,\mathrm{do}(X_2))$ *from* $P(X_1,X_2,X_3)$ *via the mass function from above.*

Example 5.3.29 (Counter example when mass functions are not strictly positive)**.** *Consider the DAG* $G = (V,E,L)$ *from Figure 12 with* $V = \{v_1,v_2,v_3\}$, $E = \{v_1 \longrightarrow v_2, v_1 \longrightarrow v_3, v_2 \longrightarrow v_3\}$. *We want to determine if we can identify* $P(X_3|\,\mathrm{do}(X_2))$ *from* $P(X_1,X_2,X_3)$ *in case we do NOT have a strictly positive mass function:* $p(x_1,x_2,x_3) > 0$. *We assume* $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_3 = \{0,1\}$.

$$\begin{aligned} p(x_1 = 1) &:= \frac{1}{2}, & p(x_2|\,\mathrm{do}(x_1)) &:= \delta_{x_1}(x_2), \\ p(x_3 = 1|\,\mathrm{do}(x_1=0,x_2=0)) &:= \frac{1}{4}, & p(x_3 = 1|\,\mathrm{do}(x_1=1,x_2=0)) &:= \frac{3}{4}, \\ p(x_3 = 1|\,\mathrm{do}(x_1=0,x_2=1)) &:= \frac{1}{8}, & p(x_3 = 1|\,\mathrm{do}(x_1=1,x_2=1)) &:= \frac{3}{8}. \end{aligned}$$

Then we get for the (X_1, X_2)-marginal of the observational distribution $P(X_1, X_2) = P(X_2|\operatorname{do}(X_1)) \otimes P(X_1)$ the mass functions:

$$\begin{aligned} p(x_1 = 0, x_2 = 0) = p(x_1 = 1, x_2 = 1) &= \frac{1}{2}, \\ p(x_1 = 1, x_2 = 0) = p(x_1 = 0, x_2 = 1) &= 0. \end{aligned}$$

Note that this shows that $P(X_1, X_2)$ and thus $P(X_1, X_2, X_3)$ do not have a strictly positive mass functions. With this a valid conditional for the observational joint distribution $P(X_1, X_2, X_3)$ conditioned on (X_1, X_2) is:

$$\begin{aligned} p(x_3 = 1|x_1 = 0, x_2 = 0) &:= \frac{1}{4}, & p(x_3 = 1|x_1 = 1, x_2 = 0) &:= \frac{3}{8}, \\ p(x_3 = 1|x_1 = 0, x_2 = 1) &:= \frac{5}{8}, & p(x_3 = 1|x_1 = 1, x_2 = 1) &:= \frac{3}{8}. \end{aligned}$$

The interventional distribution is given by:

$$\begin{aligned} p(x_3|\operatorname{do}(x_2)) &= \sum_{x_1} p(x_3|\operatorname{do}(x_1, x_2)) \cdot p(x_1), \\ p(x_3 = 1|\operatorname{do}(x_2 = 0)) &= \frac{1}{2}\left(p(x_3 = 1|\operatorname{do}(x_1 = 0, x_2 = 0)) + p(x_3 = 1|\operatorname{do}(x_1 = 1, x_2 = 0))\right) \\ &= \frac{1}{2}\left(\frac{1}{4} + \frac{3}{4}\right) = \frac{1}{2}, \\ p(x_3 = 1|\operatorname{do}(x_2 = 1)) &= \frac{1}{2}\left(p(x_3 = 1|\operatorname{do}(x_1 = 0, x_2 = 1)) + p(x_3 = 1|\operatorname{do}(x_1 = 1, x_2 = 1))\right) \\ &= \frac{1}{2}\left(\frac{1}{8} + \frac{3}{8}\right) = \frac{1}{4}. \end{aligned}$$

On the other hand, using the other conditional mass functions instead, gives us:

$$\begin{aligned} \hat{p}(x_3|x_2) &:= \sum_{x_1} p(x_3|x_1, x_2) \cdot p(x_1), \\ \hat{p}(x_3 = 1|x_2 = 0) &= \frac{1}{2}\left(p(x_3 = 1|x_1 = 0, x_2 = 0) + p(x_3 = 1|x_1 = 1, x_2 = 0)\right) \\ &= \frac{1}{2}\left(\frac{1}{4} + \frac{1}{2}\right) = \frac{3}{8} \\ &\neq \frac{1}{2} = p(x_3 = 1|\operatorname{do}(x_2 = 0)), \\ \hat{p}(x_3 = 1|x_2 = 1) &= \frac{1}{2}\left(p(x_3 = 1|x_1 = 0, x_2 = 1) + p(x_3 = 1|x_1 = 1, x_2 = 1)\right) \\ &= \frac{1}{2}\left(\frac{5}{8} + \frac{1}{8}\right) = \frac{3}{8} \\ &\neq \frac{1}{4} = p(x_3 = 1|\operatorname{do}(x_2 = 1)). \end{aligned}$$

This shows that the "surrogate" Markov kernel $\hat{P}(X_3|X_2) := P(X_3|X_1, X_2) \circ P(X_1)$, which would be proposed by both the ID-algorithm and the backdoor criterion, is NOT equal to the interventional Markov kernel $P(X_3|\operatorname{do}(X_2)) = P(X_3|\operatorname{do}(X_1, X_2)) \circ P(X_1)$, not even $P(X_2)$-almost-surely.

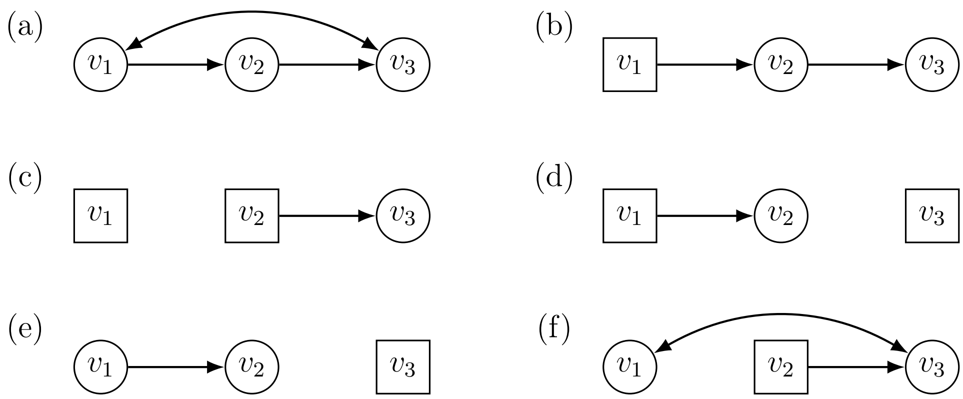


Figure 13: An ADMG and its mutilations, corresponding to the interventional Markov kernels: (a) $\mathcal{Q}[\{v_1, v_2, v_3\}]$, (b) $\mathcal{Q}[\{v_2, v_3\}]$, (c) $\mathcal{Q}[\{v_3\}]$, (d) $\mathcal{Q}[\{v_2\}]$, (e) $\mathcal{Q}[\{v_1, v_2\}]$, (f) $\mathcal{Q}[\{v_2, v_3\}]$.

Example 5.3.30. *Consider the ADMG $G = (V, E, L)$ from Figure 13 with $V = \{v_1, v_2, v_3\}$, $E = \{v_1 \longrightarrow v_2, v_2 \longrightarrow v_3\}$, $J = \emptyset$ and $L = \{v_1 \leftrightarrow v_3\}$. We want to determine if we can identify $P(X_3|\operatorname{do}(X_1))$ from $P(X_1, X_2, X_3)$ in case we have a discrete CBN with strictly positive mass function: $p(x_1, x_2, x_3) > 0$. For this let $B := \{v_3\}$ and $C := \{v_1\}$. Note that we have the topological order $v_1 < v_2 < v_3$. We then follow the steps of the ID-algorithm:*

1. $\mathcal{Q}[V](x_1, x_2, x_3) := p(x_1, x_2, x_3)$.
2. $B^C = \operatorname{Anc}^{G_{\operatorname{do}(C)}}(B) \setminus C = \{v_2, v_3\}$.
3. $\mathcal{D}[B^C] = \{S_1 = \{v_3\}, S_2 = \{v_2\}\}$, *see Figure 13 (b).*
4. *For $S_1 = \{v_3\}$:*
   a) $D_{1,0} = \operatorname{Dist}^{[V]}(S_1) = \{v_1, v_3\}$, *see Figure 13 (a), (f). Compute:*

$$
\begin{aligned}
q(x_1) &= \mathcal{Q}[V](x_1) && = p(x_1), \\
q(x_3|x_1, x_2) &= \frac{\mathcal{Q}[V](x_1, x_2, x_3)}{\mathcal{Q}[V](x_1, x_2)} && = p(x_3|x_1, x_2), \\
\mathcal{Q}[D_{1,0}](x_1, x_3|x_2) &= q(x_3|x_1, x_2) \cdot q(x_1) && = p(x_3|x_1, x_2) \cdot p(x_1).
\end{aligned}
$$

   b) $A_{1,1} = \operatorname{Anc}^{[D_{1,0}]}(S_1) = \{v_3\}$. *see Figure 13 (f), (c). Compute:*

$$
\mathcal{Q}[A_{1,1}](x_3|x_2) = \sum_{x_1} \mathcal{Q}[D_{1,0}](x_1, x_3|x_2) = \sum_{x_1} p(x_3|x_1, x_2) \cdot p(x_1).
$$

c) $D_{1,1} = \mathrm{Dist}^{[A_{1,1}]}(S_1) = \{v_3\} = A_{1,1}$, *thus* $\breve{S}_1 = A_{1,1}$. *Compute:*

$$\mathcal{Q}[\breve{S}_1](x_3|x_2) = \mathcal{Q}[A_{1,1}](x_3|x_2) = \sum_{x_1} p(x_3|x_1, x_2) \cdot p(x_1).$$

d) $S_1 = \{v_3\} = \breve{S}_1$. *Compute:*

$$\mathcal{Q}[S_1](x_3|x_2) = \mathcal{Q}[\breve{S}_1](x_3|x_2) = \sum_{x_1} p(x_3|x_1, x_2) \cdot p(x_1).$$

5. *For* $S_2 = \{v_2\}$*:*

   a) $D_{2,0} = \mathrm{Dist}^{[V]}(S_2) = \{v_2\}$, *see Figure 13 (a), (d). Compute:*

$$\mathcal{Q}[D_{2,0}](x_2|x_1) = q(x_2|x_1) = \frac{\mathcal{Q}[V](x_1, x_2)}{\mathcal{Q}[V](x_1)} = p(x_2|x_1).$$

   b) $A_{2,1} = \mathrm{Anc}^{[D_{2,0}]}(S_1) = \{v_2\} = D_{2,0}$. *thus* $\breve{S}_2 = D_{2,0}$. *Compute:*

$$\mathcal{Q}[\breve{S}_2](x_2|x_1) = \mathcal{Q}[D_{2,0}](x_2|x_1) = p(x_2|x_1).$$

   c) $S_2 = \{v_2\} = \breve{S}_2$. *Compute:*

$$\mathcal{Q}[S_2](x_2|x_1) = \mathcal{Q}[\breve{S}_2](x_2|x_1) = p(x_2|x_1).$$

6. *Since both* $\breve{S}_1 = S_1 = \{v_3\}$ *and* $\breve{S}_2 = S_2 = \{v_2\}$ *we can compute:*

$$\begin{aligned}
\mathcal{Q}[B^C](x_2, x_3|x_1) &= \mathcal{Q}[S_2](x_2|x_1) \cdot \mathcal{Q}[S_1](x_3|x_2) \\
&= p(x_2|x_1) \cdot \sum_{x_1'} p(x_3|x_1', x_2) \cdot p(x_1'), \\
p(x_3|\,\mathrm{do}(x_1)) &= \sum_{x_2} \mathcal{Q}[B^C](x_2, x_3|x_1) \\
&= \sum_{x_2} p(x_2|x_1) \cdot \sum_{x_1'} p(x_3|x_1', x_2) \cdot p(x_1') \\
&= \sum_{x_1', x_2'} p(x_3|x_1', x_2') \cdot p(x_1') \cdot p(x_2'|x_1).
\end{aligned}$$

So we can identify $P(X_3|\,\mathrm{do}(X_1))$ *from* $P(X_1, X_2, X_3)$ *via the mass function from above.*

Proofs - Soundness Criteria We have seen in Corollary 5.3.18 that the *ID-Algorithm* 5.3.17 is *sound up to oracle choices*. In this subsection we want to investigate the possibility of other forms of soundness that would allow for stronger forms of identifiability and/or trackability.

Theorem 5.3.31 (Soundness up to null-sets)**.** *Let $G = (J, V, E, L)$ be a CADMG with a fixed topological order $<$. Consider the class of L-CBNs M with observable CADMG G such that the following holds:*

1. *The measurable spaces $\mathcal{X}_v$ come equipped with a measure μ_v, $v \in V$,*
2. *for every subset $D \subseteq V$ the interventional Markov kernel $\mathcal{Q}[D] = P(X_D|\,\mathrm{do}(X_{J\cup V\setminus D}))$ is **absolute continuous** w.r.t. the product measure $\mu_D := \otimes_{v\in D}\mu_v$ and vice versa:*

$$\mu_D \ll \mathcal{Q}[D] \ll \mu_D.$$

If the ID-algorithm does not produce FAIL for input $B, C \subseteq G$, then $P(X_B|\,\mathrm{do}(X_{J\cup C}))$ is "almost-surely" trackable from $P(X_V|\,\mathrm{do}(X_J))$ and G for such CBNs M, i.e. the Markov kernel that was output by the ID-algorithm equals $P(X_B|\,\mathrm{do}(X_{J\cup C}))$ up to a $\mu_{V\setminus B^C}$-null set in $\mathcal{X}_{J\cup V\setminus B^C}$.

Proof. Since by the second assumption we have for each μ_v and any fixed value $x_{\{v\}^c}$:

$$P(X_v|\,\mathrm{do}(X_{\{v\}^c} = x_{\{v\}^c})) \ll \mu_v \ll P(X_v|\,\mathrm{do}(X_{\{v\}^c} = x_{\{v\}^c})),$$

we can w.l.o.g. assume that the μ_v are probability measures for $v \in V$.

a) Now consider any subset $A \subseteq V$ and $D \in \mathcal{D}[A]$ and $v \in D$. We abbreviate:

$$G' := G_{\mathrm{do}(V\setminus A)}, \qquad A_< := \mathrm{Pred}_<^{[A]}(v), \qquad D_< := \mathrm{Pred}_<^{[D]}(v).$$

Assume that we have a Markov kernel:

$$\mathcal{K}[A] = \tilde{P}(X_A|\,\mathrm{do}(X_{J\cup V\setminus A})) : \mathcal{X}_{J\cup V\setminus A} \dashrightarrow \mathcal{X}_A,$$

such that:

$$\mathcal{K}[A] = \mathcal{Q}[A] \qquad \mu_{V\setminus A}\text{-a.s.}$$

Note that the almost sure equality from above implies the equality:

$$\mathcal{K}[A] \otimes \mu_{V\setminus A} = \mathcal{Q}[A] \otimes \mu_{V\setminus A}. \tag{34}$$

We want to show that if we perfom the steps of the ID-algorithm that computes $\mathcal{Q}[D]$ from $\mathcal{Q}[A]$ on $\mathcal{K}[A]$ then the corresponding output, abbreviated as $\mathcal{K}[D]$, satisfies:

$$\mathcal{K}[D] = \mathcal{Q}[D] \qquad \mu_{V\setminus D}\text{-a.s.}$$

For this consider any version of the conditional of the following marginal of $\mathcal{K}[A]$:

$$\tilde{P}(X_{A_\le}|\,\mathrm{do}(X_{J\cup V\setminus A})) \qquad \text{w.r.t.} \qquad \tilde{P}(X_{A_<}|\,\mathrm{do}(X_{J\cup V\setminus A})),$$

which we will denote by:

$$K(X_v|X_{J\cup V\setminus A_\geq}).$$

This by definition will satisfy:

$$\tilde{P}(X_{A_\leq}|\,\mathrm{do}(X_{J\cup V\setminus A})) = K(X_v|X_{J\cup V\setminus A_\geq}) \otimes \tilde{P}(X_{A_<}|\,\mathrm{do}(X_{J\cup V\setminus A})). \tag{35}$$

This then implies:

$$\begin{aligned}
&P(X_{A_\leq}|\,\mathrm{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A})\\
&\overset{\text{Eq. 34}}{=} \tilde{P}(X_{A_\leq}|\,\mathrm{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A})\\
&\overset{\text{Eq. 35}}{=} K(X_v|X_{J\cup V\setminus A_\geq}) \otimes \tilde{P}(X_{A_<}|\,\mathrm{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A})\\
&\overset{\text{Eq. 34}}{=} K(X_v|X_{J\cup V\setminus A_\geq}) \otimes P(X_{A_<}|\,\mathrm{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A}).
\end{aligned}$$

This shows that $K(X_v|X_{J\cup V\setminus A_\geq})$ is a version of the conditional of:

$$P(X_{A_\leq}|\,\mathrm{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A}) \qquad \text{w.r.t.} \qquad P(X_{A_<}|\,\mathrm{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A}).$$

Note that by the interventional ordered local Markov property, Proposition 5.3.10, there exists a Markov kernel $Q(X_v|X_{\mathrm{Mb}^{G'}_<(v)})$ that simultaneuosly is a version of both:

$$P(X_v|X_{A_<}, \mathrm{do}(X_{J\cup V\setminus A})) \qquad \text{and} \qquad P(X_v|X_{D_<}, \mathrm{do}(X_{J\cup V\setminus D})),$$

and is thus, in particular, another version of the conditional of:

$$P(X_{A_\leq}|\,\mathrm{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A}) \qquad \text{w.r.t.} \qquad P(X_{A_<}|\,\mathrm{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A}).$$

Now consider the (measurable) set where these two conditional Markov kernels deviate:

$$\tilde{N} := \left\{ x_{J\cup V\setminus A_\geq} \in \mathcal{X}_{J\cup V\setminus A_\geq} \,\middle|\, K(X_v|X_{J\cup V\setminus A_\geq} = x_{J\cup V\setminus A_\geq}) \neq Q(X_v|X_{\mathrm{Mb}^{G'}_<(v)} = x_{\mathrm{Mb}^{G'}_<(v)}) \right\},$$

and $N := \tilde{N} \times \mathcal{X}_{A_\geq} \subseteq \mathcal{X}_{J\cup V}$. Since conditional Markov kernels are essentially unique we get that for every $x_J \in \mathcal{X}_J$ we have:

$$\left(\mathcal{Q}[A] \otimes \mu_{V\setminus A}\right)(N^{x_J}|x_J) = 0.$$

Since, by assumption, we have: $\mu_A \ll \mathcal{Q}[A]$, we get for every $x_J \in \mathcal{X}_J$:

$$\left(\mu_D \otimes \mu_{V\setminus D}\right)(N^{x_J}) = \mu_V(N^{x_J}) = \left(\mu_A \otimes \mu_{V\setminus A}\right)(N^{x_J}) = 0.$$

Since, by assumption, we also have: $\mathcal{Q}[D] \ll \mu_D$, we get for every $x_J \in \mathcal{X}_J$:

$$\left(\mathcal{Q}[D] \otimes \mu_{V\setminus D}\right)(N^{x_J}|x_J) = 0.$$

Let $\hat{N} := \tilde{N} \times \mathcal{X}_{A_\geq \setminus D_\geq}$. Since $D_\geq \subseteq A_\geq$ the set N is of the form:

$$N = \tilde{N} \times \mathcal{X}_{A_\geq} = \hat{N} \times \mathcal{X}_{D_\geq}.$$

So the above shows that we have for every $x_J \in \mathcal{X}_J$:

$$\big(P(X_{D_<}|\operatorname{do}(X_{J\cup V\setminus D}))\otimes \mu_{V\setminus D}(X_{V\setminus D})\big)\,(\hat{N}^{x_J}|x_J) = \big(\mathcal{Q}[D]\otimes\mu_{V\setminus D}\big)\,(N^{x_J}|x_J) = 0.$$

This shows that $K(X_v|X_{J\cup V\setminus A_\geq})$ and $Q(X_v|X_{\mathrm{Mb}^{G'}_{<}(v)})$ agree up to a measurable $\big(P(X_{D_<}|\operatorname{do}(X_{J\cup V\setminus D}))\otimes\mu_{V\setminus D}(X_{V\setminus D})\big)$-null set. Remember that $Q(X_v|X_{\mathrm{Mb}^{G'}_{<}(v)})$ satisfies:

$$P(X_{D_\leq}|\operatorname{do}(X_{J\cup V\setminus D})) = Q(X_v|X_{\mathrm{Mb}^{G'}_{<}(v)})\otimes P(X_{D_<}|\operatorname{do}(X_{J\cup V\setminus D})).$$

Together with the above we thus get:

$$\begin{aligned}
&P(X_{D_\leq}|\operatorname{do}(X_{J\cup V\setminus D}))\otimes\mu_{V\setminus D}(X_{V\setminus D})\\
&= Q(X_v|X_{\mathrm{Mb}^{G'}_{<}(v)})\otimes P(X_{D_<}|\operatorname{do}(X_{J\cup V\setminus D}))\otimes\mu_{V\setminus D}(X_{V\setminus D})\\
&= K(X_v|X_{J\cup V\setminus A_\geq})\otimes P(X_{D_<}|\operatorname{do}(X_{J\cup V\setminus D}))\otimes\mu_{V\setminus D}(X_{V\setminus D}).
\end{aligned}$$

This shows that $K(X_v|X_{J\cup V\setminus A_\geq})$ is version of the conditional of:

$$P(X_{D_\leq}|\operatorname{do}(X_{J\cup V\setminus D}))\otimes\mu_{V\setminus D}(X_{V\setminus D}) \qquad \text{w.r.t.} \qquad P(X_{D_<}|\operatorname{do}(X_{J\cup V\setminus D}))\otimes\mu_{V\setminus D}(X_{V\setminus D}).$$

If we let v run through $D=\{v_1,\dots,v_K\}$, $v_1<v_2<\cdots<v_K$ in reverse topological order we inductively get:

$$\begin{aligned}
&\mathcal{Q}[D]\otimes\mu_{V\setminus D}\\
&= P(X_D|\operatorname{do}(X_{J\cup V\setminus D}))\otimes\mu_{V\setminus D}(X_{V\setminus D})\\
&= K(X_{v_K}|X_{J\cup V\setminus A_{\geq v_K}})\otimes P(X_{D_{<v_K}}|\operatorname{do}(X_{J\cup V\setminus D}))\otimes\mu_{V\setminus D}(X_{V\setminus D})\\
&= K(X_{v_K}|X_{J\cup V\setminus A_{\geq v_K}})\otimes P(X_{D_{\leq v_{K-1}}}|\operatorname{do}(X_{J\cup V\setminus D}))\otimes\mu_{V\setminus D}(X_{V\setminus D})\\
&= K(X_{v_K}|X_{J\cup V\setminus A_{\geq v_K}})\otimes K(X_{v_{K-1}}|X_{J\cup V\setminus A_{\geq v_{K-1}}})\otimes P(X_{D_{<v_{K-1}}}|\operatorname{do}(X_{J\cup V\setminus D}))\otimes\mu_{V\setminus D}(X_{V\setminus D})\\
&= \cdots\\
&= \left(\bigotimes_{v\in D}^{>} K(X_v|X_{J\cup V\setminus A_{\geq v}})\right)\otimes\mu_{V\setminus D}(X_{V\setminus D}).
\end{aligned}$$

Since such factorizations are essentially unique we get that:

$$\mathcal{Q}[D] = \bigotimes_{v\in D}^{>} K(X_v|X_{J\cup V\setminus A_{\geq v}}) =: \mathcal{K}[D] \qquad \mu_{V\setminus D}\text{-a.s.}$$

This shows the claim.

b) We now reverse the situation. For a subset $A\subseteq V$ and every $D\in\mathcal{D}[A]=\{D_1,\dots,D_L\}$, consider that we are given a Markov kernel:

$$\mathcal{K}[D] = \tilde{P}(X_D|\operatorname{do}(X_{J\cup V\setminus D})) : \mathcal{X}_{J\cup V\setminus D} \dashrightarrow \mathcal{X}_D,$$

such that:

$$\mathcal{K}[D] = \mathcal{Q}[D] \qquad \mu_{V\setminus D}\text{-a.s.},$$

which implies the equality:

$$\mathcal{K}[D] \otimes \mu_{V\setminus D} = \mathcal{Q}[D] \otimes \mu_{V\setminus D}. \tag{36}$$

We want to show that we then also have:

$$\mathcal{K}[A] := \left[\bigotimes_{D\in\mathcal{D}[A]}^{>}\right] \mathcal{K}[D] = \mathcal{Q}[A] \qquad \mu_{V\setminus A}\text{-a.s.}$$

For this fix a node $v \in D$ and note that $Q(X_v|X_{\mathrm{Mb}_<^{G'}(v)})$ satisfies:

$$P(X_{D_\leq}|\,\mathrm{do}(X_{J\cup V\setminus D})) = Q(X_v|X_{\mathrm{Mb}_<^{G'}(v)}) \otimes P(X_{D_<}|\,\mathrm{do}(X_{J\cup V\setminus D})). \tag{37}$$

We then get the equalities:

$$\begin{aligned}
&\tilde{P}(X_{D_\leq}|\,\mathrm{do}(X_{J\cup V\setminus D})) \otimes \mu_{V\setminus D}(X_{V\setminus D}) \\
&\overset{\text{Eq. 36}}{=} P(X_{D_\leq}|\,\mathrm{do}(X_{J\cup V\setminus D})) \otimes \mu_{V\setminus D}(X_{V\setminus D}) \\
&\overset{\text{Eq. 37}}{=} Q(X_v|X_{\mathrm{Mb}_<^{G'}(v)}) \otimes P(X_{D_<}|\,\mathrm{do}(X_{J\cup V\setminus D})) \otimes \mu_{V\setminus D}(X_{V\setminus D}) \\
&\overset{\text{Eq. 36}}{=} Q(X_v|X_{\mathrm{Mb}_<^{G'}(v)}) \otimes \tilde{P}(X_{D_<}|\,\mathrm{do}(X_{J\cup V\setminus D})) \otimes \mu_{V\setminus D}(X_{V\setminus D}).
\end{aligned}$$

So $Q(X_v|X_{\mathrm{Mb}_<^{G'}(v)})$ is a conditional of:

$$\tilde{P}(X_{D_\leq}|\,\mathrm{do}(X_{J\cup V\setminus D})) \otimes \mu_{V\setminus D}(X_{V\setminus D}) \qquad \text{w.r.t.} \qquad \tilde{P}(X_{D_<}|\,\mathrm{do}(X_{J\cup V\setminus D})) \otimes \mu_{V\setminus D}(X_{V\setminus D}),$$

that does not depend on $X_{A_\geq}$ (as $\mathrm{Mb}_<^{G'}(v) \subseteq A_<$).

Now consider any other version of the conditional of:

$$\tilde{P}(X_{D_\leq}|\,\mathrm{do}(X_{J\cup V\setminus D})) \otimes \mu_{V\setminus D}(X_{V\setminus D}) \qquad \text{w.r.t.} \qquad \tilde{P}(X_{D_<}|\,\mathrm{do}(X_{J\cup V\setminus D})) \otimes \mu_{V\setminus D}(X_{V\setminus D}),$$

that does not depend on variables attached to $A_\geq$ and which we will denote by:

$$K(X_v|X_{J\cup V\setminus A_\geq}).$$

Note that such a Markov kernel exists, as $Q(X_v|X_{\mathrm{Mb}_<^{G'}(v)})$ is such one.

The same argumentation with $K(X_v|X_{J\cup V\setminus A_\geq})$ in place of $Q(X_v|X_{\mathrm{Mb}_<^{G'}(v)})$, using Eq. 36, shows that both, $Q(X_v|X_{\mathrm{Mb}_<^{G'}(v)})$ and $K(X_v|X_{J\cup V\setminus A_\geq})$, are then conditionals of

$$P(X_{D_\leq}|\,\mathrm{do}(X_{J\cup V\setminus D})) \otimes \mu_{V\setminus D}(X_{V\setminus D}) \qquad \text{w.r.t.} \qquad P(X_{D_<}|\,\mathrm{do}(X_{J\cup V\setminus D})) \otimes \mu_{V\setminus D}(X_{V\setminus D}),$$

that do not depend on $X_{A_\geq}$. Now let:

$$\tilde{N} := \left\{ x_{J\cup V\setminus A_\geq} \in \mathcal{X}_{J\cup V\setminus A_\geq} \,\middle|\, K(X_v|X_{J\cup V\setminus A_\geq} = x_{J\cup V\setminus A_\geq}) \neq Q(X_v|X_{\mathrm{Mb}_<^{G'}(v)} = x_{\mathrm{Mb}_<^{G}(v)}) \right\},$$

and $N := \tilde{N} \times \mathcal{X}_{A_\geq} \subseteq \mathcal{X}_{J\cup V}$. Again, since both are versions of the same conditional we get:

$$\left(\mathcal{Q}[D] \otimes \mu_{V\setminus D}\right)(N^{x_J}|x_J) = 0,$$

for every $x_J \in \mathcal{X}_J$. Since $\mathcal{Q}[D] \ll \mu_D$ we get for every $x_J \in \mathcal{X}_J$:

$$\left(\mu_A \otimes \mu_{V\setminus A}\right)(N^{x_J}) = \mu_V(N^{x_J}) = \left(\mu_D \otimes \mu_{V\setminus D}\right)(N^{x_J}) = 0.$$

Since also $\mathcal{Q}[A] \ll \mu_A$ we get for every $x_J \in \mathcal{X}_J$:

$$\left(\mathcal{Q}[A] \otimes \mu_{V\setminus A}\right)(N^{x_J}|x_J) = 0.$$

Since $N = \tilde{N} \times \mathcal{X}_{A_\geq}$ we get for every $x_J \in \mathcal{X}_J$:

$$\left(P(X_{A_<}|\operatorname{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A})\right)(\tilde{N}^{x_J}|x_J) = \left(\mathcal{Q}[A] \otimes \mu_{V\setminus A}\right)(N^{x_J}|x_J) = 0.$$

This shows that the Markov kernels $Q(X_v|X_{\mathrm{Mb}^{\mathcal{G}'}_<(v)})$ and $K(X_v|X_{J\cup V\setminus A_\geq})$ are equal up to some $\left(P(X_{A_<}|\operatorname{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A})\right)$-null set. Note that with this we get the factorization, using $A = \{v_1, \dots, v_K\}$, $v_1 < \cdots < v_K$:

$$\begin{aligned}
&\mathcal{Q}[A] \otimes \mu_{V\setminus A} \\
&= P(X_{A_{\leq v_K}}|\operatorname{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A}) \\
&= Q(X_{v_K}|X_{\mathrm{Mb}^{\mathcal{G}'}_<(v_K)}) \otimes P(X_{A_{<v_K}}|\operatorname{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A}) \\
&= K(X_{v_K}|X_{J\cup V\setminus A_{\geq v_K}}) \otimes P(X_{A_{<v_K}}|\operatorname{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A}) \\
&= K(X_{v_K}|X_{J\cup V\setminus A_{\geq v_K}}) \otimes P(X_{A_{\leq v_{K-1}}}|\operatorname{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A}) \\
&= K(X_{v_K}|X_{J\cup V\setminus A_{\geq v_K}}) \otimes Q(X_{v_K}|X_{\mathrm{Mb}^{\mathcal{G}'}_<(v_K)}) \otimes P(X_{A_{<v_{K-1}}}|\operatorname{do}(X_{J\cup V\setminus A})) \otimes \mu_{V\setminus A}(X_{V\setminus A}) \\
&= \cdots \\
&= \left(\bigotimes_{v\in A}^{>} K(X_v|X_{J\cup V\setminus A_\geq})\right) \otimes \mu_{V\setminus A}(X_{V\setminus A}) \\
&= \left(\left[\bigotimes_{D\in\mathcal{D}[A]}^{>}\right] \mathcal{K}[D]\right) \otimes \mu_{V\setminus A}(X_{V\setminus A}) \\
&= \mathcal{K}[A] \otimes \mu_{V\setminus A}(X_{V\setminus A}).
\end{aligned}$$

Since such factorizations are essentially unique we get:

$$\mathcal{K}[A] = \mathcal{Q}[A] \qquad \mu_{V\setminus A}\text{-a.s.}$$

This shows the claim.

c) Now let $D \subseteq V$ and $A \subseteq D$ with $A = \mathrm{Anc}^{[D]}(A)$. Consider that we are given a Markov kernel:

$$\mathcal{K}[D] = \tilde{P}(X_D|\operatorname{do}(X_{J\cup V\setminus D})) : \mathcal{X}_{J\cup V\setminus D} \dashrightarrow \mathcal{X}_D,$$

such that:

$$\mathcal{K}[D] = \mathcal{Q}[D] \qquad \mu_{V\setminus D}\text{-a.s.},$$

which implies the equality:

$$\mathcal{K}[D] \otimes \mu_{V\setminus D} = \mathcal{Q}[D] \otimes \mu_{V\setminus D}.$$

We want to show that the A-marginal of $\mathcal{K}[D]$ equals $\mathcal{Q}[A]$ up to $\mu_{V\setminus A}$-null set.

For this let $\mathcal{K}[A]$ be the A-marginal of $\mathcal{K}[D]$:

$$\mathcal{K}[A] := \tilde{P}(X_A|\,\mathrm{do}(X_{J\cup V\setminus D})) : \mathcal{X}_{J\cup V\setminus A} \to \mathcal{X}_{J\cup V\setminus D} \dashrightarrow \mathcal{X}_A.$$

Note that $\mathcal{Q}[A]$ is the A-marginal of $\mathcal{Q}[D]$. Marginalizing out $X_{D\setminus A}$ on both sides in the above equation gives us:

$$\mathcal{K}[A] \otimes \mu_{V\setminus D} = \mathcal{Q}[A] \otimes \mu_{V\setminus D}.$$

Multiplying both sides with $\mu_{D\setminus A}$ gives:

$$\mathcal{K}[A] \otimes \mu_{V\setminus A} = \mathcal{K}[A] \otimes \mu_{V\setminus D} \otimes \mu_{D\setminus A} = \mathcal{Q}[A] \otimes \mu_{V\setminus D} \otimes \mu_{D\setminus A} = \mathcal{Q}[A] \otimes \mu_{V\setminus A}.$$

Since such factorizations are essentially unique we get:

$$\mathcal{K}[A] = \mathcal{Q}[A] \qquad \mu_{V\setminus A}\text{-a.s.}$$

This shows the claim.

This covers all cases of the ID-algorithm and thus shows the claim. □

Theorem 5.3.32 (Soundness up to continuous choices for strictly positive CBNs)**.** *Let $G = (J, V, E, L)$ be a CADMG with a fixed topological order $<$. Consider the class of L-CBNs M with observable CADMG G such that the following holds:*

1. *The spaces $\mathcal{X}_v$ are Polish spaces for $v \in J \cup V$,*
2. *for every subset $D \subseteq V$ the interventional Markov kernel $\mathcal{Q}[D] = P(X_D|\,\mathrm{do}(X_{J\cup V\setminus D}))$ is **strictly positive** (on non-empty open subsets of $\mathcal{X}_D$), and:*
3. *for every $v \in D$ the Markov kernel $Q(X_v|X_{\mathrm{Mb}_<^{G_{\mathrm{do}(D^c)}}(v)})$ can be chosen to be **continuous**, viewed as a map: $\mathcal{X}_{\mathrm{Mb}_<^{G_{\mathrm{do}(D^c)}}(v)} \to \mathcal{P}(\mathcal{X}_v)$.*

If the ID-algorithm does not produce FAIL for input $B, C \subseteq G$, then $P(X_B|\,\mathrm{do}(X_{J\cup C}))$ is identifiable and trackable "up to continuous choices of conditional Markov kernels" from $P(X_V|\,\mathrm{do}(X_J))$ and G for such CBNs M, i.e. if every occuring conditional Markov kernel is chosen to be continuous (which will always be possible by the assumptions made).

Proof. For a district $D \in \mathcal{D}[V]$ and $v \in D$, by assumption, there exists a *continuous* version of $Q(X_v|X_{\mathrm{Mb}_<^G(v)})$, which is also a version of:

$$P(X_v|X_{\mathrm{Pred}_<^{[V]}(v)}, \mathrm{do}(X_J)) \qquad \text{and} \qquad P(X_v|X_{\mathrm{Pred}_<^{[D]}(v)}, \mathrm{do}(X_{J\cup V\setminus D})).$$

We abbreviate $V_< := \mathrm{Pred}_<^{[V]}(v)$ and $D_< := \mathrm{Pred}_<^{[D]}(v)$ in the following.

Now consider any *continuous* version of the conditional Markov kernel $P(X_v|X_{V_<}, \mathrm{do}(X_J))$. Note that such a version always exists because $Q(X_v|X_{\mathrm{Mb}_<^{\mathcal{G}}(v)})$ is already an existing continuous version. Since $P(X_{V_<}|\,\mathrm{do}(X_J))$ is strictly positive, as the marginal of $\mathcal{Q}[V]$, Lemma 2.4.23 implies then the "sure" equality:

$$P(X_v|X_{V_<}, \mathrm{do}(X_J) = Q(X_v|X_{\mathrm{Mb}_<^{\mathcal{G}}(v)}).$$

So any continuous version of $P(X_v|X_{V_<}, \mathrm{do}(X_J))$ necessarily agrees with $Q(X_v|X_{\mathrm{Mb}_<^{\mathcal{G}}(v)})$ on all points.

Similarly, using the same arguments, we get that $Q(X_v|X_{\mathrm{Mb}_<^{\mathcal{G}}(v)})$ is "surely" equal to every *continuous* version of the conditional $P(X_v|X_{D_<}, \mathrm{do}(X_{J\cup V\setminus D}))$, as also the Markov kernel $P(X_{D_<}|\,\mathrm{do}(X_{J\cup V\setminus D}))$ is strictly positive, as a marginal of $\mathcal{Q}[D]$. So we get:

$$Q(X_v|X_{\mathrm{Mb}_<^{\mathcal{G}}(v)}) = P(X_v|X_{D_<}, \mathrm{do}(X_{J\cup V\setminus D})).$$

This means that if we pick a/the *continuous* version of the conditional $P(X_v|X_{V_<}, \mathrm{do}(X_J))$ then it is "surely" equal to the/every *continuous* version of the conditional $P(X_v|X_{D_<}, \mathrm{do}(X_{J\cup V\setminus D}))$:

$$P(X_v|X_{V_<}, \mathrm{do}(X_J)) = Q(X_v|X_{\mathrm{Mb}_<^{\mathcal{G}}(v)}) = P(X_v|X_{D_<}, \mathrm{do}(X_{J\cup V\setminus D})).$$

These arguments, repeated for subgraphs, then show that in the ID-algorithm for every occuring conditional and product (e.g. for districts and the final product) we end up with distinct and correct choices for all Markov kernels. This then also shows the identifiability of such CBNs M (in the not-FAIL case). □

Lemma 5.3.33. *Consider an L-CBN:*

$$M = \left(G^+ = (J, (V, U), E^+), \left(P_v(X_v|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)}))\right)_{v\in V\cup U}\right),$$

with observable CADMG $G = (J, V, E, L)$ and fixed topological order $<$. Assume that for every $v \in V$ we have a measure μ_v on $\mathcal{X}_v$ such that $P_v(X_v|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)}))$ has a (strictly positive) density w.r.t. μ_v:

$$p(x_v|\,\mathrm{do}(x_{\mathrm{Pa}^{G^+}(v)})) > 0.$$

Furthermore, we put for $x_V \in \mathcal{X}_V$, $x_U \in \mathcal{X}_U$, $x_J \in \mathcal{X}_J$:

$$p(x_V|x_U, \mathrm{do}(x_J)) := \prod_{v\in V} p(x_v|\,\mathrm{do}(x_{\mathrm{Pa}^{G^+}(v)})),$$

and then integrate in reverse order of $<$:

$$p(x_V|\,\mathrm{do}(x_J)) := \int \cdots \int_{\mathcal{X}_U} p(x_V|x_U, \mathrm{do}(x_J)) \bigotimes_{u\in U}^{>} P_u(X_u \in dx_u|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)} = x_{\mathrm{Pa}^{G^+}(v)})).$$

Then the former is a (strictly positive) density of $P(X_V, X_U|\,\mathrm{do}(X_J))$ w.r.t.

$$\bigotimes_{v\in V}^{>} \mu_v \otimes^{>} \bigotimes_{u\in U}^{>} P_u(X_u|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)})),$$

and the latter a (strictly positive) density of $P(X_V|\,\mathrm{do}(X_J))$ w.r.t.

$$\mu_V := \bigotimes_{v \in V} \mu_v.$$

Similarly, for every $D \subseteq V$ the interventional Markov kernel $P(X_D|\,\mathrm{do}(X_{J \cup V \setminus D}))$ has a (strictly positive) density w.r.t. $\mu_D := \bigotimes_{v \in D} \mu_v$.

Proof. The claim can be shown by integrating the above densities over product sets $A = \prod_v A_v$. Inductively we can use Fubini's theorem and:

$$\int_{A_v} p(x_v|\,\mathrm{do}(x_{\mathrm{Pa}^{G^+}(v)}))\,\mu_v(dx_v) = P_v(X_v \in A_v|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)} = x_{\mathrm{Pa}^{G^+}(v)})).$$

Regarding strict positivity, note that if $f(x) > 0$ for all x, then $\int f\,d\mu > 0$ for non-trivial μ. So strict positivity is preserved through integration. $\square$

6. Structural Causal Models (with Inputs)

Structural Causal Models (SCMs), also known as Non-Parametric Structural Equation Models (NP-SEMs), provide a class of causal models that can model causal cycles. SCMs trace back to the early work on path analysis by geneticist Sewall Wright [Wri21], made their way to econometrics [Wri28, Haa43, SW60] and the social sciences [Bol89] under the name Structural Equation Models (SEMs), and became popular in AI due to the work of Judea Pearl [Pea09] and many others. In these lecture notes, we give a modern treatment inspired by our own research on the matter [BFPM21, FM20]. We introduce here a more general class of causal models, namely SCMs 'with inputs' (briefly: 'iSCMs').

6.1. Motivation

While causal Bayesian networks (with input nodes and latent variables) provide an expressive causal modeling class, there is an important aspect of causality that cannot be modeled with causal Bayesian networks, namely causal *cycles*. For example, increasing temperature at the poles may cause sea ice to melt, which leads to more absorption of sunlight because white ice is replaced by blue sea water, which in turn leads to further temperature increase (see also Figure 14(a)). Because a causal Bayesian network is acyclic by definition, such a model can only be described by a causal Bayesian network by introducing multiple variables corresponding with measurements of the same quantities at different points in time (Figure 14(b)). In contrast, an SCM can directly represent causal cycles and is often appropriate for modeling systems with fast feedback processes that are stable, i.e., where negative feedback dominates potential positive feedback. An illustrative example is a system composed of different masses connected via springs in an environment with friction (see also Section 6.12).

Another example of the limitation of Causal Bayesian Networks is the following, which shows that even if a 'fine-grained' causal model is acyclic, merging variables may introduce cycles at a more 'coarse-grained' level of description.

Example 6.1.1. *Suppose that we have a CBN with the ADMG in Figure 15(a), representing four variables X_1, X_2, X_3, X_4. If we chose an alternative representation in terms of pairs $X_{13} := (X_1, X_3)$ and $X_{24} := (X_2, X_4)$, then we would end up with a CBN with the DMG in Figure 15(b). However, that is a contradiction as the graph of a CBN is acyclic by definition.*

This example shows that the class of CBNs is not closed under the operation of merging variables. The class of (simple) SCMs with inputs to be introduced later is actually closed under the operation of merging variables.

Finally, there exist systems in which the directionality of causal relations is context-dependent.

Example 6.1.2. *Consider Ohm's law $V = IR$ (voltage equals current times resistance) to model the voltage across and current through a resistance. If we connect the resistance to a voltage source, the voltage determines the current. If we connect the resistance to a*

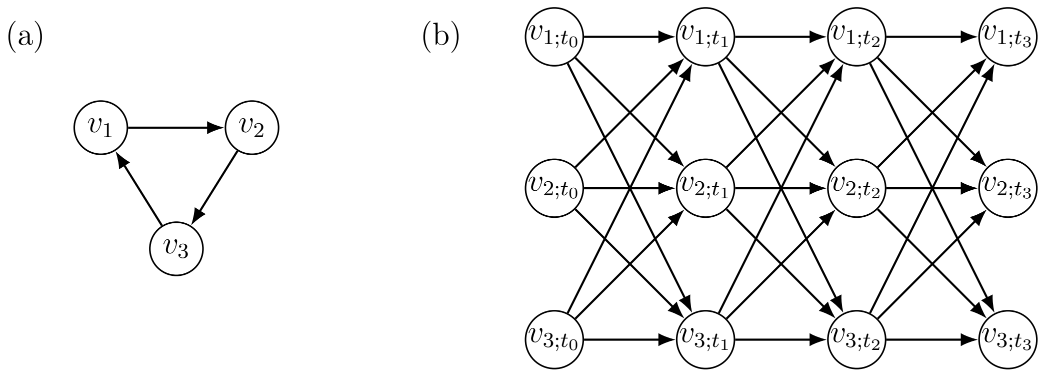


Figure 14: (a) Directed Graph (DG) representing a causal cycle. As an example, v_1 could be the average temperature in a certain area at the North pole, v_2 the amount of sea ice present in the area, and v_3 the amount of sunlight absorbed in the area. This gives an example of a positive (self-reinforcing) feedback loop. (b) Alternative Directed Acyclic Graph (DAG) where the variables correspond with the same quantities but measured at different time points $t_0 < t_1 < t_2 < t_3$.

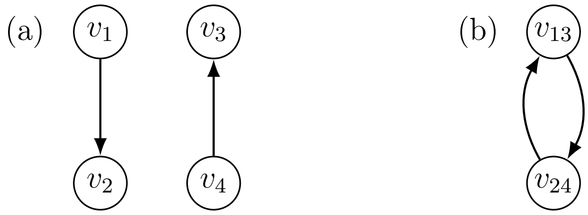


Figure 15: (a) ADMG with output nodes v_1, v_2, v_3, v_4 corresponding with endogenous variables X_1, X_2, X_3, X_4. (b) DMG corresponding to a coarser representation obtained by merging variables into $X_{13} := (X_1, X_3)$ and $X_{24} := (X_2, X_4)$.

current source, then it is the other way around: the current determines the voltage. Both cases separately can be modeled with a CBN (Figure 16(a–b), respectively). If we let a coin flip determine which of the two sources the resistance is connected to, we obtain a mixture which cannot be modeled as a single CBN (Figure 16(c)).

Similar behavior is often encountered in complex systems in biology, chemistry, engineering and economy. This is yet another motivation to extend the causal modeling framework to allow for cycles.

In this chapter, we will introduce the class of SCMs with inputs, which generalize CBNs (with latent variables and input nodes) to allow for cycles, and allows us to deal elegantly with all motivating examples discussed here. In addition, SCMs (with inputs) support counterfactual reasoning, another capability that CBNs lack.

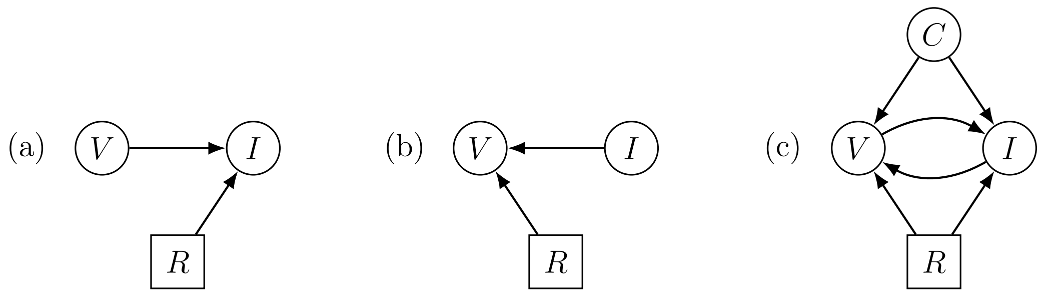


Figure 16: Different causal models corresponding to modeling the current through a resistance using Ohm's law. (a) Voltage causes current. (b) Current causes voltage. (c) Mixture model where the causal relationship between voltage and current depends on the result of a coin flip.

6.2. Definition

An SCM is specified in terms of (deterministic) functions and distributions, rather than in terms of Markov kernels ('stochastic functions'). Conceptually, in an SCM the focus lies on modeling the *causal mechanisms* as deterministic functions at an abstract level. Such a function expresses the value of an effect variable in terms of its direct cause variables. We assume that the joint value of the direct cause variables determines a unique value of the effect variable, so that the value of the effect is *completely* explained by the values of its direct causes. To prevent an infinite regress (where the values of the cause variables must be explained by the values of their direct causes, and so on *ad infinitum*), the model does not specify a causal mechanism for *all* variables that occur in the model. This divides the variables into two types: the 'endogenous' variables, each of which has an associated causal mechanism that expresses its value in terms of its direct causes, and the 'exogenous' variables, for which no such causal mechanisms are specified. We can distinguish two types of exogenous variables: the 'exogenous inputs', for which the model only describes the range of their possible joint values, and the 'exogenous random' variables for which the model in addition describes their probability distribution.

In contrast with many definitions encountered in the literature,[29] we here explicitly distinguish three types of variables: endogenous variables, exogenous latent random variables and exogenous observed input variables. We will refer to those objects as 'SCMs with inputs' (or 'iSCMs').

Definition 6.2.1 (Structural Causal Model with Inputs)**.** *A **Structural Causal Model with Inputs** (iSCM) is a tuple $M = (J, V, W, \mathcal{X}, P, f)$ such that*

- *J, V, W are disjoint finite sets of labels for the **exogenous input variables**, the **endogenous variables** and the **exogenous random variables**, respectively;*
- *the **domain** $\mathcal{X} = \prod_{i \in J \dot\cup V \dot\cup W} \mathcal{X}_i$ is a product of standard measurable spaces $\mathcal{X}_i$;*

[29]For example, [Pea09] only formally distinguishes exogenous latent random variables and endogenous variables.

- *the **exogenous distribution** P is a probability distribution on $\mathcal{X}_W$ that factorizes as a product $P = \bigotimes_{w \in W} P_w$ of probability distributions $P_w \in \mathcal{P}(\mathcal{X}_w)$;*[30]
- *the **causal mechanisms** are specified by the measurable function $f : \mathcal{X} \to \mathcal{X}_V$.*

*For the special case $J = \emptyset$ (no exogenous inputs) we refer to the tuple $M = (V, W, \mathcal{X}, P, f)$ as a **Structural Causal Model (SCM)**.*

The endogenous variables are the variables whose causal relations we wish to model, whereas the exogenous variables are required to model the remaining causes of the endogenous variables.

Definition 6.2.2 (Parameterized iSCM)**.** *Often, the causal mechanism f and the exogenous distribution P depend (measurably) on **exogenous parameters** $\theta \in \mathcal{X}_\Theta$, which we may make explicit by writing f^θ and P^θ instead, giving a parameterized iSCM $M^\theta = \big(J, V, W, \mathcal{X}, P^\theta, f^\theta\big)$. The family $(M^\theta)_{\theta \in \mathcal{X}_\Theta}$ is then an iSCM family.*[31]

One can also think about an iSCM as describing an input/output system, with free inputs J, random inputs W with distribution P, outputs V and 'modular' input/output mechanisms f_v for $v \in V$. Structural causal models can be regarded as a marriage of statistical models as traditionally used in statistics (a parameterized family of distributions) with deterministic causal models (deterministic input/output systems) that are used informally in disciplines like physics and engineering.

Remark 6.2.3. *There are four crucial assumptions embodied in the modeling approach using iSCMs:*

1. *Exogenous variables (i.e., exogenous input variables, exogenous random variables, and exogenous parameters) are **not caused** by endogenous variables.*
2. *Exogenous variables are **variation independent**: their joint range is the Cartesian product of the range of each exogenous variable.*
3. *Exogenous random variables are **probabilistically independent**, and their distribution is independent of the exogenous input variables (but may depend on exogenous parameters).*
4. *Exogenous parameters describe **population** properties, while exogenous random and input variables describe **individual** quantities. This means that if we are modeling a population of N 'units' that are all subject to the same causal mechanism, we can make N copies of the iSCM, one for each individual, where the copies*

[30]Because of the heavy use of the symbol 'P' we will often use the notation 'P_M' to refer to the exogenous distribution of iSCM M.

[31]In line with the convention in machine learning, the word 'model' refers to a iSCM with a fixed choice of the parameters, and 'model family' to a family of models indexed measurably by parameters. This contrasts with the terminology in statistics, where a family of distributions indexed measurably by parameters is called a 'statistical model'.

share the same causal mechanism f and the same exogenous distribution $P(X_W)$, and in the parametric case, the same parameter θ. A frequentist statistician might treat exogenous parameters as 'global' exogenous input variables, while a Bayesian statistician might treat exogenous parameters as 'global' exogenous random variables by also specifying a prior $P(\theta) \in \mathcal{P}(\mathcal{X}_\Theta)$.

Remark 6.2.4. *Not all types of variables need to be present. Rather than giving separate definitions for 'degenerate' cases, we can stay in the formalism by defining what happens for empty label sets. For example, suppose the iSCM is deterministic, i.e., $W = \emptyset$. Then $\mathcal{X}_W$ is an empty product (i.e., a product over 0 spaces), and by definition becomes a space $\ast = \{*\}$ with a single element $*$, with the trivial sigma algebra $\{\emptyset, \{*\}\}$. The only possible probability distribution on such a space is the trivial distribution, i.e., $P(\{*\}) = 1$. Similarly, SCMs (without inputs) can be treated as the special case where $J = \emptyset$.*

6.3. Solving the Structural Equations

6.3.1. Structural Equations and Potential Outcomes

An iSCM defines *structural equations*, which are used to define the *potential outcomes* of the iSCM.

Definition 6.3.1 (Potential outcomes, structural equations)**.** *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM and $x_J \in \mathcal{X}_J$ an input value. A random variable $X_{V \cup W}^{\mathrm{do}(x_J)}$ with codomain $\mathcal{X}_V \times \mathcal{X}_W$ is called a **potential outcome of** M **for input** x_J if the following two conditions hold:*[32]

1. *its W-component has the exogenous distribution specified by M:*

$$X_W^{\mathrm{do}(x_J)} \sim P,$$

2. *it satisfies the **structural equations entailed by** M **for input** x_J:*

$$X_V^{\mathrm{do}(x_J)} = f\big(x_J, X_V^{\mathrm{do}(x_J)}, X_W^{\mathrm{do}(x_J)}\big) \text{ a.s..} \tag{38}$$

*In case $J = \emptyset$ we also write $X_{V \cup W} := X_{V \cup W}^{\mathrm{do}(*)}$ and refer to it simply as an **outcome of** M.*

The iSCM encodes the probability distributions of its (potential) outcomes. However, the structural equations (38) may have no solution or may have multiple different solutions. Therefore, even if a (potential) outcome exists (for a given input), it could be that its distribution is not uniquely determined by the iSCM.

[32] Another notation for potential outcomes, commonly encountered in the literature, is $X_{V \cup W}(x_J)$.

Example 6.3.2. *Consider an iSCM with parameters $\alpha, \beta \in \mathbb{R}$, endogenous real-valued variables X_1, X_2, real-valued exogenous input X_3, and structural equations for input $x_3 \in \mathbb{R}$ given by:*

$$\begin{cases} X_1^{\mathrm{do}(x_3)} = \alpha X_2^{\mathrm{do}(x_3)} \\ X_2^{\mathrm{do}(x_3)} = \beta X_1^{\mathrm{do}(x_3)} + x_3. \end{cases}$$

If $\alpha\beta = 1$, then $(X_1^{\mathrm{do}(x_3=0)}, X_2^{\mathrm{do}(x_3=0)}) = (\alpha x, x)$ is a potential outcome for input $x_3 = 0$ for any value of $x \in \mathbb{R}$. Any mixture of these potential outcomes is also a potential outcome for input $x_3 = 0$. If $\alpha\beta = 1$, then for input $x_3 \neq 0$, the iSCM admits no potential outcomes. If $\alpha\beta \neq 1$, then the potential outcomes are unique and given by $(X_1^{\mathrm{do}(x_3)}, X_2^{\mathrm{do}(x_3)}) = (\frac{\alpha x_3}{1-\alpha\beta}, \frac{x_3}{1-\alpha\beta})$.

In practice, an iSCM is often specified more informally by writing down the corresponding structural equations and by giving the exogenous distribution of X_W. Any variables appearing on the r.h.s. of some structural equation that do not correspond with a structural equation for which that variable appears on the l.h.s., nor have a specified distribution, are then implicitly taken as exogenous inputs (or parameters).

Example 6.3.3 (Linear regression model with fixed design for the effect in terms of its cause (population level))**.** *Assume that $Y = \alpha X + \beta + \epsilon$ with $Y \in \mathbb{R}$ representing the effect, $X \in \mathbb{R}$ the cause, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ independent normally distributed measurement noise, and $\alpha, \beta \in \mathbb{R}$, $\sigma^2 \in (0, \infty)$ parameters.*

For a fixed parameter value, this can be understood as the specification of an iSCM $M^{(\alpha,\beta,\sigma^2)} = \left(J, V, W, \mathcal{X}, P^{\sigma^2}, f^{\alpha,\beta}\right)$ with $J = \{X\}$, $V = \{Y\}$, $W = \{\epsilon\}$, $\mathcal{X} = \mathbb{R}^3$, exogenous distribution $P^{\sigma^2} = \mathcal{N}(0, \sigma^2)$ and causal mechanism $f^{\alpha,\beta} : \mathbb{R}^3 \to \mathbb{R} : (x, y, \epsilon) \mapsto \alpha x + \beta + \epsilon$.

If $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $x \in \mathbb{R}$, then $(Y^{\mathrm{do}(x)}, \epsilon^{\mathrm{do}(x)}) := (\alpha x + \beta + \epsilon, \epsilon)$ is a potential outcome of $M^{(\alpha,\beta,\sigma^2)}$ for input x. Often, one just refers to the component $Y^{\mathrm{do}(x)}$ as the potential outcome.

One often is interested in modeling a 'population' consisting of 'individuals' (for example, all the patients in a medical trial). In that case, the iSCM is often considered a 'template' where the full causal model on all individuals is obtained by making a copy of the template for each individual.

Example 6.3.4 (Linear regression model with fixed design for the effect in terms of its cause (individual level))**.** *The 'individual-level' iSCM for a population consisting of N individuals corresponding to the 'population-level' iSCM of Example 6.3.3 has:*

- *Exogenous input variable $J = \{X_1, \dots, X_N\}$;*
- *Exogenous random variable indices $W = \{\epsilon_1, \dots, \epsilon_N\}$;*
- *Endogenous variable indices $V = \{Y_1, \dots, Y_N\}$;*
- *Joint state space $\mathcal{X} = \prod_{i=1}^N \mathcal{X}_i = \prod_{i=1}^N (\mathbb{R} \times \mathbb{R} \times \mathbb{R}) = \mathbb{R}^{3N}$;*

- *Parameters* $(\alpha, \beta, \sigma^2) \in \mathbb{R}^2 \times (0, \infty)$*;*
- *Exogenous distribution* $P^\sigma(X_W) = \bigotimes_{i=1}^N P^\sigma(X_{W_i})$ *with* $P^\sigma(X_{W_i}) \sim \mathcal{N}(0, \sigma^2)$*;*
- *Causal mechanisms*
$$f_{y_i}^{\alpha,\beta} : (x_i, y_i, \epsilon_i) \mapsto \alpha x_i + \beta + \epsilon_i$$
for $i = 1, \dots, N$.

Note how the parameters are 'shared' across individuals, but the other variables are not. We get potential outcomes with Y*-component* $Y^{\mathrm{do}(x)} = (Y_1^{\mathrm{do}(x_1,\dots,x_N)}, \dots, Y_n^{\mathrm{do}(x_1,\dots,x_N)})$*. This individual-level model assumes that all individuals are independent entities that do not interact.*[33]

6.3.2. Solutions

The language of conditional random variables allows us to give a neat definition of the *solution* of an iSCM, which can also be thought of as a measurable family of (potential) outcomes with a shared underlying probability space.

Definition 6.3.5 (Solutions of an iSCM)**.** *Let* $M = (J, V, W, \mathcal{X}, P, f)$ *be an iSCM. Let* $(\mathcal{U} \times \mathcal{X}_J, K(U|X_J))$ *be a transition probability space and* $X : \mathcal{U} \times \mathcal{X}_J \to \mathcal{X}$ *be a conditional random variable. If its marginal (push-forward) Markov kernel* $K(X_W|X_J)$ *is constant in* X_J *and coincides with the exogenous distribution specified by* M*:*

$$\forall x_J \in X_J : \qquad K(X_W|x_J) \sim P,$$

and X *satisfies the* ***structural equations*** *entailed by* M*:*

$$X_V = f(X_J, X_V, X_W) \qquad a.s., \tag{39}$$

then X *is called a* ***solution of*** M*.*[34] *Since the* X_J *component is trivial, we also refer to the component*

$$X_{V\cup W} : \mathcal{U} \times \mathcal{X}_J \to \mathcal{X}_{V \cup W}$$

as a solution of M*.*

Remark 6.3.6. *Let* $M = (J, V, W, \mathcal{X}, P, f)$ *be an iSCM and* $X : \mathcal{U} \times \mathcal{X}_J \to \mathcal{X}_{V\cup W}$ *be a solution of* M*. Then for any* $x_J \in \mathcal{X}_J$*,*

$$X_{V\cup W}^{\mathrm{do}(x_J)} : \mathcal{U} \to \mathcal{X}_{V\cup W} : u \mapsto X(u, x_J)$$

is a potential outcome of M *for input* x_J*. Its distribution coincides with* $K(X_{V\cup W}|X_J = x_J)$*.*

[33] Of course, that is not always a realistic assumption, for example when studying the efficacy of vaccines, vaccinating one individual could not only protect that particular individual from contracting a disease, but also other individuals in its vicinity.

[34] We write 'a.s.' as a shorthand for $K(U|X_J)$-a.s., extending the corresponding convention for random variables.

The converse does not hold: specifying a complete family of (potential) outcomes $(X_{V\cup W}^{\mathrm{do}(x_J)})_{x_J\in\mathcal{X}_J}$ does not necessarily give a solution, as the family may not be measurable in x_J.

Not every iSCM has solutions. Also, if they exist, solutions are not necessarily unique, even if they have the same underlying transition probability space.

Example 6.3.7. *Consider an SCM with parameters $\alpha, \beta, \mu \in \mathbb{R}$, $\sigma^2 \in [0, \infty)$, endogenous real-valued variables X_1, X_2, exogenous random real-valued variable W_1 with exogenous distribution $\mathcal{N}(\mu, \sigma^2)$, and structural equations*

$$\begin{cases} X_1 = \alpha X_2 \\ X_2 = \beta X_1 + W_1. \end{cases}$$

If $\alpha\beta = 1$, $\mu = 0$ and $\sigma^2 = 0$, then $(X_1, X_2, W_1) = (\alpha x, x, W_1)$ is a solution for any $x \in \mathbb{R}$ and $W_1 \sim \mathcal{N}(\mu, \sigma^2)$. Any mixture of those solutions is also a solution in that case. If $\alpha\beta = 1$ and $\mu \neq 0$ or $\sigma^2 \neq 0$, then the SCM admits no solutions. If $\alpha\beta \neq 1$, then all solutions (X_1, X_2, W_1) satisfy $(X_1, X_2) = (\frac{\alpha W_1}{1-\alpha\beta}, \frac{W_1}{1-\alpha\beta})$ a.s. and $W_1 \sim \mathcal{N}(\mu, \sigma^2)$.

The following remark relates the terminology to the cases most often considered in the literature.

Remark 6.3.8. *If $J = \emptyset$, a solution of an SCM can be identified with a random variable $X_{V\cup W}$ with codomain $\mathcal{X}_V \times \mathcal{X}_W$ such that $X_W \sim P$ and that satisfies the structural equations:*

$$X_v = f_v\big(X_V, X_W\big) \text{ a.s.} \tag{40}$$

for each $v \in V$.

6.3.3. Markov Kernels of Solutions

Each solution of an iSCM 'has' a Markov kernel (similarly to how each random variable has a distribution).

Notation 6.3.9. *Let conditional random variable $X : \mathcal{U} \times \mathcal{X}_J \to \mathcal{X}$ on transition probability space $(\mathcal{U} \times \mathcal{X}_J, K(U|X_J))$ be a solution of an iSCM $M = (J, V, W, \mathcal{X}, P, f)$. Its push-forward*

$$P(X \mid \mathrm{do}(X_J)) := K(X|X_J) = X_* K(U|X_J)$$

is a Markov kernel $\mathcal{X}_J \dashrightarrow \mathcal{X}$ that we refer to as ***the Markov kernel of** M **corresponding to** X*. Since the J-component is trivial, we also refer to its marginal $P(X_{V\cup W} \mid \mathrm{do}(X_J))$ as such.*

Not all solutions of an iSCM may yield the same Markov kernel (see also Example 6.3.7). Therefore, even if the iSCM M is specified, the notation '$P(X \mid \mathrm{do}(X_J))$' is ambiguous, since it does not specify which solution the Markov kernel comes from. Because we will mostly restrict attention to so-called 'simple' iSCMs for which the Markov kernel turns out to be unique, we will not worry about this.

Remark 6.3.10. *In case $J = \emptyset$ for an SCM M, the Markov kernel $P(X \mid \mathrm{do}(X_J))$ corresponding to a solution X can be identified with its distribution $P(X)$, and one often refers to it as the* ***distribution of*** *M* ***corresponding to*** *X.*

6.3.4. Null sets (again)

We will frequently need to specify that certain properties or identities hold 'almost surely' in a specific sense. This is analogous to the notion of 'almost surely' in probability theory, but because we are dealing with parameterized families of probability distributions, the notion becomes more intricate. We will use it in the theory of iSCMs in the same way that the notion of $P(X_W | \mathrm{do}(X_J))$-null sets are used in the theory of CBNs (see Definition 2.3.15).

In the following definitions, an important subtlety to be aware of is that the ordering of the quantifiers matters, and reordering the quantifiers changes the meaning of the statement. The specific ordering was chosen to correspond with the data generating processes that are modeled via a sampling process as described in Section 6.3.6.

Definition 6.3.11 (iSCM Null Sets)**.** *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM.*

1. *We call a set $\tilde{N} \subseteq \mathcal{X}_J \times \mathcal{X}_W$ an M-**null set** if for each $x_J \in \mathcal{X}_J$, the section $\tilde{N}_{x_J} = \{x_W \in \mathcal{X}_W : (x_J, x_W) \in \tilde{N}\}$ is a $P_M(X_W)$-null set.*

2. *We call a set $\tilde{N} \subseteq \mathcal{X}_J \times \mathcal{X}_W$ a **measurable** M-**null set** if $\tilde{N} \in \mathcal{B}_J \otimes \mathcal{B}_W$ and $\tilde{N}$ is an M-null set.*

3. *We call a set $N \subseteq \mathcal{X}_J \times \mathcal{X}_W \times \mathcal{X}_V$ a **(measurable)** M-**null set** if it is contained in some $\tilde{N} \times \mathcal{X}_V$ with $\tilde{N}$ a (measurable) M-null set.*

4. *More generally, for $L \subseteq V$, we call a set $N \subseteq \mathcal{X}_J \times \mathcal{X}_W \times \mathcal{X}_L$ a **(measurable)** M-**null set** if it is contained in some $\tilde{N} \times \mathcal{X}_L$ with $\tilde{N}$ a (measurable) M-null set.*

Note that the first notion corresponds with that of a '$P_M(X_W | \mathrm{do}(X_J))$-null set' (cf. Definition 2.3.15). It also corresponds with that of a '$P_M(X_W)$-null set in $\mathcal{X}_{J \cup W}$' (cf. Remark 5.1.7). Note that subsets of M-null sets are also M-null sets, and countable unions of M-null sets are also M-null sets.

Definition 6.3.12 ('Almost surely' according to an iSCM)**.** *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM.*

1. *Let $\tilde{\pi} : \mathcal{X}_J \times \mathcal{X}_W \to \{0, 1\}$ be a binary property. We say that $\tilde{\pi}$ **holds** M-**a.s.** if*

$$\{(x_J, x_W) \in \mathcal{X}_J \times \mathcal{X}_W : \tilde{\pi}(x_J, x_W) = 0\}$$

   is contained in a M-null set.

2. *Let $\pi : \mathcal{X}_J \times \mathcal{X}_W \times \mathcal{X}_V \to \{0, 1\}$ be a binary property. We say that π **holds** M-**a.s.** if:*

$$\{(x_J, x_W, x_V) \in \mathcal{X}_J \times \mathcal{X}_W \times \mathcal{X}_V : \pi(x_J, x_W, x_V) = 0\}$$

   is contained in a M-null set.

We can use this e.g. to define M-a.s. equality of sets / functions / Markov kernels.

Note that both notions are compatible if we consider $\tilde{\pi}(x_J, x_W) := \forall x_V \in \mathcal{X}_V : \pi(x_J, x_W, x_V)$. We can think about the statement '$\pi(x_J, x_W, x_V)$ holds M-a.s.' as the claim that for all $x_J \in \mathcal{X}_J$, for $P_M(X_W)$-almost all $x_W \in \mathcal{X}_W$, for all $x_V \in \mathcal{X}_V$, $\pi(x_J, x_W, x_V)$ holds. The ordering of the quantifiers matters: the 'for all x_V' is the innermost quantifier, and the 'for all x_J' is the outermost quantifier.

The following lemma translates an M-a.s. property to an a.s. property of a random variable.

Lemma 6.3.13. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. Let $\pi : \mathcal{X}_J \times \mathcal{X}_W \times \mathcal{X}_V \to \{0, 1\}$ be a binary property. Let $x_J \in \mathcal{X}_J$. Let $X_{V \cup W}$ be a random variable with values in $\mathcal{X}_V \times \mathcal{X}_W$ and such that its W-marginal is distributed according to the exogenous distribution of M (that is, $X_W \sim P_M$). Then:*

$$\pi(x_J, x_W, x_V) \quad M\text{-a.s.} \qquad \Longrightarrow \qquad \pi(x_J, X_W, X_V) \quad \text{a.s.}.$$

Proof. Since π holds M-a.s., for the fixed $x_J \in \mathcal{X}_J$ this means that for P_M-almost all x_W:

$$\forall x_V \in \mathcal{X}_V : \pi(x_J, x_W, x_V) = 1.$$

Since $X_W \sim P_M$, we get that:

$$[\forall x_V \in \mathcal{X}_V : \pi(x_J, X_W, x_V) = 1] \quad P(X_W)\text{-a.s.}.$$

Hence:

$$[\forall x_V \in \mathcal{X}_V : \pi(x_J, X_W, x_V) = 1] \quad \text{a.s.},$$

which implies:

$$\pi(x_J, X_W, X_V) = 1 \quad \text{a.s.}$$

□

We can do something analogous for a conditional random variable.

Lemma 6.3.14. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. Let $\pi : \mathcal{X}_J \times \mathcal{X}_W \times \mathcal{X}_V \to \{0, 1\}$ be a binary property. Let $(\mathcal{U} \times \mathcal{X}_J, K(U|X_J))$ be a transition probability space and $X : \mathcal{U} \times \mathcal{X}_J \to \mathcal{X}$ be a conditional random variable such that its marginal (push-forward) Markov kernel $K(X_W|X_J)$ is constant and coincides with the exogenous distribution of M (that is, $K(X_W|X_J) = P_M$). Then:*

$$\pi(x_J, x_W, x_V) \quad M\text{-a.s.} \qquad \Longrightarrow \qquad \pi(X_J, X_W, X_V) \quad \text{a.s.}.$$

Proof. Since π holds M-a.s., for all $x_J \in \mathcal{X}_J$, for P_M-almost all x_W:

$$\forall x_V \in \mathcal{X}_V : \pi(x_J, x_W, x_V) = 1.$$

Since $K(X_W|X_J) = P_M$, we get that:

$$[\forall x_V \in \mathcal{X}_V : \pi(X_J, X_W, x_V) = 1] \quad K(X_W|X_J)\text{-a.s.}.$$

Hence:

$$[\forall x_V \in \mathcal{X}_V : \pi(X_J, X_W, x_V) = 1] \quad \text{a.s.},$$

which implies:

$$\pi(X_J, X_W, X_V) = 1 \quad \text{a.s.}.$$

□

6.3.5. (Partial) Solution functions

A fundamental notion in the theory of (cyclic) iSCMs is that of its 'solution functions'. A solution function is a function that maps an exogenous state to a corresponding solution of the structural equations, in a measurable way. A technical subtlety is that it suffices to specify solution functions up to a certain type of null set.

Definition 6.3.15 (Solution function of an iSCM)**.** *Given an iSCM $M = (J, V, W, \mathcal{X}, P, f)$, we call a function $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ a* ***solution function of*** *M if it is measurable and*

$$g(x_J, x_W) = f\big(x_J, g(x_J, x_W), x_W\big) \quad M\text{-a.s.}$$

We call M ***solvable*** *if it has a solution function.*

Later on we will need a more refined notion of solvability, which essentially amounts to solving a *subset* of the structural equations.

Definition 6.3.16 (Partial solution function)**.** *Given an iSCM $M = (J, V, W, \mathcal{X}, P, f)$ and a nonempty subset $L \subseteq V$, we call a function $g^{[L]} : \mathcal{X}_J \times \mathcal{X}_{V \setminus L} \times \mathcal{X}_W \to \mathcal{X}_L$ a* ***(partial) solution function of*** *M* ***w.r.t.*** *L if it is measurable and*

$$g^{[L]}(x_J, x_{V\setminus L}, x_W) = f_L\big(x_J, x_{V\setminus L}, g^{[L]}(x_J, x_{V\setminus L}, x_W), x_W\big) \quad M\text{-a.s.}$$

We call M ***solvable w.r.t.*** *L if it has a partial solution function w.r.t. L.*

Note that 'M is solvable w.r.t. V' means the same as 'M is solvable'.

We can use solution functions of an iSCM to construct (potential) outcomes, solutions and Markov kernels of the iSCM.

Proposition 6.3.17. *If g is a solution function for iSCM $M = (J, V, W, \mathcal{X}, P, f)$, then for any random variable $X_W : \mathcal{U} \to \mathcal{X}_W$ with distribution P:*

1. *$X_{V,W}^{\mathrm{do}(x_J)} := (g(x_J, X_W), X_W)$ is a potential outcome for M for input $x_J \in \mathcal{X}_J$;*
2. *$X_{V,W} := (g(X_J, X_W), X_W)$ is a solution of M with underlying transition probability space $(\mathcal{U} \times \mathcal{X}_J, P(U))$;*
3. *its corresponding Markov kernel is the push-forward*

$$(g, \mathrm{id}_{\mathcal{X}_W})_*(P \otimes \delta(X_J|X_J)).$$

Proof. 1. This follows immediately with the help of Lemma 6.3.13.

2. Note first that g is measurable, and hence $X_{V,W} := (g(X_J, X_W), X_W)$ defines a conditional random variable with underlying transition probability space $(\mathcal{U} \times \mathcal{X}_J, P(U))$. The claim now follows immediately with the help of Lemma 6.3.14.

3. By definition.

$\square$

Not all (potential) outcomes, solutions and Markov kernels of an iSCM can be obtained in this way. For example, mixtures of solutions are also solutions, but not all mixtures can be obtained as the push-forward through a solution function.

Example 6.3.18. *For an iSCM with endogenous real variables X_1, X_2 and structural equations*

$$\begin{cases} X_1 &= X_2, \\ X_2 &= X_1^3, \end{cases}$$

any real-valued random variable Y for which $P(Y \in \{-1, 0, 1\}) = 1$ provides a solution $(X_1, X_2) := (Y, Y)$. This includes all mixtures over the three possible states $(-1, -1)$, $(0, 0)$, $(1, 1)$, which form a two-dimensional convex space. However, it has only three solution functions (mapping $*$ to either $(-1, -1)$, $(0, 0)$ or $(1, 1)$). Therefore, only three solutions can be constructed from a solution function as in Proposition 6.3.17 (namely the extreme points of the convex space).

How can we check in practice whether a given iSCM is (partially) solvable? The following result provides a condition that is relatively easy to check in practice. It shows that if the structural equations have (almost everywhere) unique solutions, then the measurability of the solution function follows from that of the causal mechanisms.

Proposition 6.3.19. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. Let $L \subseteq V$. Let*

$$U^{[L]} = \{x_{J \cup (V \setminus L) \cup W} \in \mathcal{X}_J \times \mathcal{X}_{V \setminus L} \times \mathcal{X}_W \mid \exists! x_L \in \mathcal{X}_L : x_L = f_L(x_J, x_{V \setminus L}, x_W, x_L)\}$$

be the set of extended exogenous states for which there exists a unique partial solution of the structural equations for L. Then $U^{[L]} = \mathcal{X}_J \times \mathcal{X}_{(V \setminus L)} \times \mathcal{X}_W$ up to a measurable M-null set if and only if there exists a measurable function $g^{[L]} : \mathcal{X}_J \times \mathcal{X}_{(V \setminus L)} \times \mathcal{X}_W \to \mathcal{X}_L$ such that

$$x_L = f_L(x) \iff x_L = g^{[L]}(x_J, x_{(V \setminus L)}, x_W) \quad M\text{-a.s.}.$$

Proof. Assume that $N = \tilde{N} \times \mathcal{X}_L$ is a measurable M-null set with $\tilde{N}^c \subseteq U^{[L]}$. We will show that there exists a measurable function $g^{[L]} : \mathcal{X}_J \times \mathcal{X}_{V \setminus L} \times \mathcal{X}_W \to \mathcal{X}_L$ s.t.

$$g^{[L]}(x_J, x_{V \setminus L}, x_W) = f_L\big(x_J, x_{V \setminus L}, x_W, g^{[L]}(x_J, x_{V \setminus L}, x_W)\big) \quad M\text{-a.s.}$$

We exploit that we are dealing with standard measurable spaces. Define the function

$$h^{[L]} : \mathcal{X}_{J \cup W \cup V} \to \mathcal{X}_L \times \mathcal{X}_L : x \mapsto (x_L, f_L(x))$$

and the diagonal $\Delta^{[L]} = \{(x_L, x_L) : x_L \in \mathcal{X}_L\} \subseteq \mathcal{X}_L^2$. Since $h^{[L]}$ is measurable and $\Delta^{[L]}$ is a Borel measurable set (because $\mathcal{X}_L$ is Hausdorff), $(h^{[L]})^{-1}(\Delta^{[L]})$ is a measurable set. Hence, $(h^{[L]})^{-1}(\Delta^{[L]}) \cap (\tilde{N}^c \times \mathcal{X}_L)$ is a measurable subset.

We now define $g_0^{[L]}$ as the function $\tilde{N}^c \mapsto \mathcal{X}_L$ such that $\mathrm{graph}(g_0^{[L]}) = (h^{[L]})^{-1}(\Delta^{[L]}) \cap \tilde{N}^c \times \mathcal{X}_L$. In other words, $g_0^{[L]}$ maps $(x_J, x_{(V\setminus L)}, x_W) \in \tilde{N}^c$ to the unique element $x_L \in \mathcal{X}_L$ such that $x_L = f_L(x_J, x_{V\setminus L}, x_W, x_L)$. By [Kec95, 14.12], because all spaces are (isomorphic to) Borel spaces, the fact that $\mathrm{graph}(g_0^{[L]})$ is a measurable set implies that $g_0^{[L]}$ is a measurable function. We can extend $g_0^{[L]}$ to a measurable function $g^{[L]} : \mathcal{X}_J \times \mathcal{X}_{(V\setminus L)} \times \mathcal{X}_W \to \mathcal{X}_L$, for example, by picking an arbitrary $x_L^0 \in \mathcal{X}_L$ and defining:

$$g^{[L]}(x_J, x_{(V\setminus L)}, x_W) = \begin{cases} g_0^{[L]}(x_J, x_{(V\setminus L)}, x_W) & (x_J, x_{(V\setminus L)}, x_W) \in \tilde{N}^c \\ x_V^0 & (x_J, x_{(V\setminus L)}, x_W) \in \tilde{N}. \end{cases}$$

It is clear that $g^{[L]}$ constructed in this way is a partial solution function of M w.r.t. L. This means:

$$x_L = f_L(x) \iff x_L = g^{[L]}(x_J, x_{(V\setminus L)}, x_W) \quad M\text{-a.s.}.$$

Conversely, since $x_{J\cup(V\setminus L)\cup W} \in U^{[L]} \quad M$-a.s.,

$$x_L = f_L(x) \implies x_L = g^{[L]}(x_J, x_{(V\setminus L)}, x_W) \quad M\text{-a.s.}.$$

Now assume that $g^{[L]} : \mathcal{X}_J \times \mathcal{X}_{(V\setminus L)} \times \mathcal{X}_W \to \mathcal{X}_L$ is a measurable function such that

$$x_L = f_L(x) \iff x_L = g^{[L]}(x_J, x_{(V\setminus L)}, x_W) \quad M\text{-a.s.}.$$

Let $A := \{x \in \mathcal{X} : x_L = f_L(x)\}$. Let $B := \{x \in \mathcal{X} : x_L = g^{[L]}(x_J, x_{(V\setminus L)}, x_W)\}$. We have already seen that $A = (h^{[L]})^{-1}(\Delta^{[L]})$ is a measurable set. Via similar reasoning (and the fact that $g^{[L]}$ is a measurable function) we can show that B is a measurable set. Hence $C := \{x \in \mathcal{X} : x_L = f_L(x) \iff x_L = g^{[L]}(x_J, x_{(V\setminus L)}, x_W)\} = (A \cap B) \cup (A^c \cap B^c)$ is a measurable set. Its complement is an M-null set by assumption. Since $C \subseteq U^{[L]}$, it follows that $U^{[L]} = \mathcal{X}_J \times \mathcal{X}_{(V\setminus L)} \times \mathcal{X}_W$ up to a measurable M-null set. □

Remark 6.3.20. *Under the weaker assumption that $(U^{[L]})^c$ is an M-null set (not necessarily measurable), it is an open question whether M is partially solvable w.r.t. L. From [Kec95, Thm 18.11], we obtain that $U^{[L]}$ is a co-analytic subset of $\mathcal{X}_J \times \mathcal{X}_{(V\setminus L)} \times \mathcal{X}_W$—that is, $(U^{[L]})^c$ is analytic—but it is not clear whether that guarantees the existence of a measurable partial solution function.*

As a more extensive example, we now formalize the chocolate-Nobel prize example discussed in Section 1.1 as different iSCMs according to some of the causal hypotheses.

Example 6.3.21. *For a given country, consider two real-valued variables: annual chocolate consumption in kilograms per capita (C), and the number of Nobel prize winners per year per capita (N). We can consider the following linear iSCM families $M^\theta = (J, V, W, \mathcal{X}, P^\theta, f^\theta)$.*

1. *N causes C: ('Nobel prizes are celebrated with massive chocolate feasts')* $J = \{N\}$, $V = \{C\}$, $W = \emptyset$, $\theta = (\alpha, \beta) \in \mathcal{X}_\Theta := \mathbb{R}^2$, $\mathcal{X}_N = \mathbb{R}$, $\mathcal{X}_C = \mathbb{R}$, $f^{\alpha,\beta} : (x_N, x_C) \mapsto \alpha + \beta x_N$. *It has structural equations*

$$X_C^{\mathrm{do}(x_N)} = \alpha + \beta x_N.$$

   For a given parameter θ, the iSCM M^θ has a unique solution function, $g^{\alpha,\beta} : x_N \mapsto \alpha + \beta x_N$, and unique potential outcomes of the form $X_C^{\mathrm{do}(x_N)} = \alpha + \beta x_N$.

2. *C causes N: ('chocolate contains brain enhancing chemicals')* $J = \{C\}$, $V = \{N\}$, $W = \emptyset$, $\theta = (\gamma, \delta) \in \mathcal{X}_\Theta := \mathbb{R}^2$, $\mathcal{X}_C = \mathbb{R}$, $\mathcal{X}_N = \mathbb{R}$, $f^{\gamma,\delta} : (x_C, x_N) \mapsto \gamma + \delta x_C$. *It has structural equations*

$$X_N^{\mathrm{do}(x_C)} = \gamma + \delta x_C.$$

   For a given parameter θ, the iSCM M^θ has a unique solution function, $g^{\gamma,\delta} : x_C \mapsto \gamma + \delta x_C$, and unique potential outcomes of the form $X_N^{\mathrm{do}(x_C)} = \gamma + \delta x_C$.

3. *W causes C and N, version 1: ('inhabitants of wealthy countries eat more chocolate and conduct more scientific research')* $J = \{W\}$, $V = \{C, N\}$, $W = \emptyset$, $\theta = (\alpha, \beta, \gamma, \delta) \in \mathcal{X}_\Theta := \mathbb{R}^4$, $\mathcal{X}_W = \mathbb{R}$, $\mathcal{X}_C = \mathbb{R}$, $\mathcal{X}_N = \mathbb{R}$, $f^\theta : (x_W, x_C, x_N) \mapsto (\alpha + \beta x_W, \gamma + \delta x_W) \in \mathcal{X}_C \times \mathcal{X}_N$. *It has structural equations*

$$X_C^{\mathrm{do}(x_W)} = \alpha + \beta x_W,$$
$$X_N^{\mathrm{do}(x_W)} = \gamma + \delta x_W.$$

   For a given parameter θ, the iSCM M^θ has unique solution function $g^\theta : x_W \mapsto (\alpha + \beta x_W, \gamma + \delta x_W) \in \mathcal{X}_C \times \mathcal{X}_N$ and it has unique potential outcomes of the form $X_{C,N}^{\mathrm{do}(x_W)} = (\alpha + \beta x_W, \gamma + \delta x_W)$.

4. *W causes C and N, version 2: ('similar to version 1, but now the probability distribution of wealth is modeled'):* $J = \emptyset$, $V = \{C, N\}$, $W = \{W\}$, $\theta = (\alpha, \beta, \gamma, \delta, \sigma) \in \mathcal{X}_\Theta := \mathbb{R}^4 \times [0, \infty)$, $\mathcal{X}_W = \mathbb{R}$, $\mathcal{X}_C = \mathbb{R}$, $\mathcal{X}_N = \mathbb{R}$, $f^\theta : (x_C, x_N, x_W) \mapsto (\alpha + \beta x_W, \gamma + \delta x_W) \in \mathcal{X}_C \times \mathcal{X}_N$, $P^\theta = \mathcal{N}(0, \sigma^2)$. *It has structural equations*

$$X_C = \alpha + \beta X_W,$$
$$X_N = \gamma + \delta X_W.$$

   For a given parameter θ, the iSCM M^θ has solution function $g^\theta : x_W \mapsto (\alpha + \beta x_W, \gamma + \delta x_W) \in \mathcal{X}_C \times \mathcal{X}_N$ and it has outcomes of the form $X_{C,N} = (\alpha + \beta X_W, \gamma + \delta X_W)$ for some random variable $X_W \sim \mathcal{N}(0, \sigma^2)$. It has many solution functions; for example, also $g^\theta + \tilde{g}\mathbb{1}_N$ for some $P(X_W)$-nullset N and arbitrary measurable function $\tilde{g} : \mathcal{X}_W \to \mathbb{R}^2$ is a solution function.

6.3.6. Sampling from an iSCM

One can use the solution function to sample from the corresponding Markov kernel of an iSCM.

Remark 6.3.22. *Given a solution function $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ of an iSCM $M = (J, V, W, \mathcal{X}, P, f)$, we can sample from its corresponding Markov kernel in the following way:*

for $i = 1, \ldots, n$ ***do***
    input $x_J^{(i)}$
    for each $w \in W$ ***do***
        sample $X_w^{(i)} \sim P(X_w)$
    end for
    for each $v \in V$ ***do***
        sample $X_v^{(i)} \leftarrow g(x_J^{(i)}, x_W^{(i)})$
    end for
    output $(x_J^{(i)}, X_W^{(i)}, X_V^{(i)})$
end for

We can think of this as a model of the data-generating process that yields a sequence of samples

$$\left(x_J^{(1)}, X_W^{(1)}, X_V^{(1)}\right), \left(x_J^{(2)}, X_W^{(2)}, X_V^{(2)}\right), \ldots, \left(x_J^{(n)}, X_W^{(n)}, X_V^{(n)}\right).$$

While the $X_W^{(i)}$ are independent, we are not assuming anything about how the values $x_J^{(i)}$ are determined. More precisely, the above sampling process will lead to the following distribution for the exogenous random variables, given a certain input sequence $x_J^{(1)}, \ldots, x_J^{(n)}$:

$$(X_W^{(1)}, \ldots, X_W^{(n)}) \mid x_J^{(1)}, \ldots, x_J^{(n)} \sim P(X_W^{(1)}) \otimes \cdots \otimes P(X_W^{(n)}).$$

In words: the $X_W^{(n)}$ are independent and identically distributed according to P, given the entire sequence $x_J^{(1)}, \ldots, x_J^{(n)}$. Jointly, the data-generating process leads to the following distribution:

$$(X_{V\cup W}^{(1)}, \ldots, X_{V\cup W}^{(n)}) \mid x_J^{(1)}, \ldots, x_J^{(n)} \sim K(X_{V\cup W}^{(1)} \,|\, x_J^{(1)}) \otimes \cdots \otimes K(X_{V\cup W}^{(n)} \,|\, x_J^{(n)}),$$

with $K(X_{V\cup W} \,|\, X_J)$ the Markov kernel obtained from the push-forward

$$(g, \mathrm{id}_{\mathcal{X}_W})_*(P \otimes \delta(X_J|X_J)).$$

Marginally on the endogenous variables, the above data-generating process will lead to the following distribution of the observed data, given a certain input sequence:

$$(X_V^{(1)}, \ldots, X_V^{(n)}) \mid x_J^{(1)}, \ldots, x_J^{(n)} \sim K(X_V^{(1)} \,|\, x_J^{(1)}) \otimes \cdots \otimes K(X_V^{(n)} \,|\, x_J^{(n)}),$$

with $K(X_V \mid X_J)$ the Markov kernel obtained from the push-forward

$$g_*(P \otimes \delta(X_J|X_J)).$$

The ordering of the operations (input, sample, calculate, output) in this sampler matters. In particular, note that the values x_J of the exogenous input variables are determined *before* the exogenous random variables X_W are sampled. Swapping this ordering may lead to dependence $X_W \not\perp\!\!\!\perp X_J$ between the exogenous random and exogenous input variables (i.e., $P(X_W \mid \mathrm{do}(X_J)) \neq P(X_W)$). The ordering chosen here is compatible with the assumption that the exogenous random variables are latent.[35] Indeed, suppose that some 'agent' decides on the values of the exogenous inputs x_J. If this agent could observe X_W before it decides on the values of the inputs, the agent could choose the values of X_J dependent on the values of X_W.[36]

If the iSCM admits multiple solution functions, this sampler depends explicitly on the choice of the solution function. So one way to think about iSCMs that admit multiple solution functions is that they are *incomplete* models of a data-generating process.

Example 6.3.23 (Continuing Example 6.3.21)**.** *We can sample from the iSCMs in Example 6.3.21 as follows:*

1. *N causes C: ('Nobel prizes are celebrated with massive chocolate feasts')*

   input α, β, n
   for $i = 1, \ldots, n$ ***do***
       input $x_N^{(i)}$;
       $X_C^{(i)} \leftarrow \alpha + \beta x_N^{(i)}$;
       output $(X_C^{(i)}, x_N^{(i)})$
   end for

2. *C causes N: ('chocolate contains brain enhancing chemicals')*

   input γ, δ, n
   for $i = 1, \ldots, n$ ***do***
       input $x_C^{(i)}$;
       $X_N^{(i)} \leftarrow \gamma + \delta x_C^{(i)}$;
       output $(x_C^{(i)}, X_N^{(i)})$
   end for

[35] One could also consider a setting in which some, but not all, information about the value x_W is accessible to the experimenter. For example, some exogenous random variables could be considered "observable" and others "latent", while they would all be considered "non-intervenable". We will not consider this as it would complicate matters further.

[36] If the iSCM is misspecified in the sense that 'nature' choses the values of X_W dependent on the values of X_J, then one can find another iSCM by reparameterizing X_W to make it independent of X_J again. Thus, the assumption that $X_W \sim P(X_W)$ rather than $X_W \sim K(X_W \mid \mathrm{do}(X_J))$ in Definition 6.2.1 can be made without loss of generality.

3. *W causes C and N, version 1: ('inhabitants of wealthy countries eat more chocolate and conduct more scientific research')*

   input $\alpha, \beta, \gamma, \delta, n$
   for $i = 1, \dots, n$ ***do***
       input $x_W^{(i)}$;
       $X_C^{(i)} \leftarrow \alpha + \beta x_W^{(i)}$;
       $X_N^{(i)} \leftarrow \gamma + \delta x_W^{(i)}$.
       output $(X_C^{(i)}, X_N^{(i)}, x_W^{(i)})$
   end for

4. *W causes C and N, version 2: ('similar to version 1, but now the probability distribution of wealth is modeled'):*

   input $\alpha, \beta, \gamma, \delta, \sigma^2, n$
   for $i = 1, \dots, n$ ***do***
       sample $X_W^{(i)} \sim \mathcal{N}(0, \sigma^2)$;
       $X_C^{(i)} \leftarrow \alpha + \beta x_W^{(i)}$;
       $X_N^{(i)} \leftarrow \gamma + \delta x_W^{(i)}$.
       output $(X_C^{(i)}, X_N^{(i)}, X_W^{(i)})$
   end for

6.4. Interventions

The reason that equations (38) are called *structural* is that one cannot simply rewrite them in the way one is used to when solving a set of equations without changing the causal semantics of the model. This can be formalized by defining how interventions affect an iSCM.

In this section we define interventions as *operations* on iSCMs that map a given iSCM and an intervention target (and optionally, an intervention value or distribution) to an intervened iSCM. The operation may change the types of the variables it is targeting. We will consider four intervention types: three variants of a hard intervention, and soft interventions. The three hard intervention variants differ in what type of variables the intervened variables become: endogenous variables, exogenous random variables, or exogenous input variables. As there are many ways in the real world to intervene on (or 'to perturb', or simply 'to change') a given system, this is only the tip of an iceberg of how one could formalize such interventions.

Remark 6.4.1. *We will* **not** *consider interventions that target exogenous random variables. This allows the modeler to make less stringent assumptions when modeling a data-generating process with an iSCM.*[37]

[37] If an intervention on an exogenous random variable in a model is really required, one can always make an endogenous copy of it, and then intervene on the endogenous copy.

6.4.1. Hard interventions

We start with hard interventions that turn all intervened variables into endogenous variables with specified values, overriding the default causal mechanisms that determined their values before the intervention was performed.

Definition 6.4.2 (Hard intervention with specified target values)**.** *Given an iSCM $M = (J, V, W, \mathcal{X}, P, f)$, an intervention target $T \subseteq J \cup V$ and an intervention value $\xi_T \in \mathcal{X}_T$, we define the* ***intervened iSCM***

$$M_{\mathrm{do}(X_T=\xi_T)} := \left(J \setminus T, V \cup T, W, \mathcal{X}, P, (f_{V\setminus T}, \xi_T)\right).$$

More explicitly, the components of the intervened causal mechanism $\tilde{f} : \mathcal{X} \to \mathcal{X}_{V\cup T}$ are given by:

$$\tilde{f}_j(x) = \begin{cases} \xi_j & j \in T \\ f_j(x) & j \in V \setminus T, \end{cases}$$

for $j \in V \cup T$.

This replaces the targeted exogenous variables by endogenous variables and adds structural equations to set their values as specified, replaces the existing structural equations of the form $X_j^{\mathrm{do}(x_J)} = f_j(x_J, X_V^{\mathrm{do}(x_J)}, X_W^{\mathrm{do}(x_J)})$ for $j \in T \cap V$ to structural equations of the simple form $X_j^{\mathrm{do}(x_{J\setminus T})} = \xi_j$, and leaves the other structural equations invariant. This operation 'endogenizes' exogenous input variables, reflecting that the intervened model now specifies their values as prescribed by the hard intervention. The values of the other endogenous variables are still determined by their original causal mechanisms.

Example 6.4.3. *A hard intervention $\mathrm{do}(X_N = \xi_N)$ changes the iSCM 'W causes C and N, version 2' from Example 6.3.21 into the iSCM with $J = \emptyset$, $V = \{C, N\}$, $W = \{W\}$, $\theta = (\alpha, \beta, \gamma, \delta, \sigma) \in \mathbb{R}^4 \times [0, \infty)$, $\mathcal{X}_W = \mathbb{R}$, $\mathcal{X}_C = \mathbb{R}$, $\mathcal{X}_N = \mathbb{R}$, $\tilde{f}^\theta : (x_C, x_N, x_W) \mapsto (\alpha + \beta x_W, \xi_N) \in \mathcal{X}_C \times \mathcal{X}_N$, $P^\theta = \mathcal{N}(0, \sigma^2)$. It has structural equations*

$$\begin{aligned} X_C &= \alpha + \beta X_W, \\ X_N &= \xi_N. \end{aligned}$$

Another common variant of hard interventions are stochastic hard interventions, where the intervention values are drawn independently from a specified (independent) distribution.

Definition 6.4.4 (Stochastic hard intervention)**.** *Given an iSCM $M = (J, V, W, \mathcal{X}, P, f)$, an intervention target $T \subseteq J \cup V$ and an intervention target distribution $Q_T \in \bigotimes_{t\in T} \mathcal{P}(\mathcal{X}_t)$, we define the* ***intervened iSCM***

$$M_{\mathrm{do}(X_T\sim Q_T)} := \left(J \setminus T, V \setminus T, W \cup T, \mathcal{X}, P \otimes Q_T, f_{V\setminus T}\right).$$

More explicitly, the intervened exogenous distribution is given by

$$P \otimes Q_T = \left[\bigotimes_{w\in W} P_w\right] \otimes \left[\bigotimes_{t\in T} Q_t\right].$$

Intuitively, this assigns random values to the intervention target variables by sampling from independent intervention distributions Q_t for $t \in T$, thereby turning the targeted variables into exogenous random variables. A hard intervention on an exogenous input variable turning it into an exogenous random variable can be interpreted as 'imposing a distribution' on the exogenous input variable. For example, if treatment is considered an exogenous input variable (the model does not specify how treatment is determined by the physician for each patient), and we then intervene to let treatment be determined by a coin flip instead (when setting up an RCT), we are imposing a distribution on the treatment variable.

Example 6.4.5. *The iSCM 'W causes C and N, version 2' from Example 6.3.21 is obtained from a stochastic intervention on the iSCM 'W causes C and N, version 1' from that example.*

The third variant of hard interventions only specifies the intervention targets, but makes no assertions about the intervention values (not even their distribution).

Definition 6.4.6 (Hard intervention with unspecified value)**.** *Given an iSCM $M = (J, V, W, \mathcal{X}, P, f)$ and an intervention target $T \subseteq J \cup V$, we define the* ***intervened iSCM***

$$M_{\mathrm{do}(T)} := \left(J \cup T, V \setminus T, W, \mathcal{X}, P, f_{V \setminus T}\right).$$

Intuitively, this operation replaces endogenous variables and exogenous random variables with exogenous input variables. The intervened model no longer specifies the causal mechanisms that determine the values of these variables, but instead treats them as exogenous inputs that are independent of the (remaining) exogenous random variables in the model. This reflects that after this hard intervention, the values for these variables are no longer determined by the system, but are set externally (e.g., by the experimenter performing the intervention) to values chosen independently of the values of the exogenous random variables, while the values of the other endogenous variables are still determined by their original causal mechanisms.

Example 6.4.7. *A hard intervention* $\mathrm{do}(N)$ *changes the iSCM 'W causes C and N, version 2' from Example 6.3.21 into the intervened iSCM with: $J = \{N\}$, $V = \{C\}$, $W = \{W\}$, $\theta = (\alpha, \beta, \gamma, \delta, \sigma) \in \mathbb{R}^4 \times [0, \infty)$, $\mathcal{X}_W = \mathbb{R}$, $\mathcal{X}_C = \mathbb{R}$, $\mathcal{X}_N = \mathbb{R}$, $\tilde{f}^\theta : (x_N, x_C, x_W) \mapsto \alpha + \beta x_W$, $P^\theta = \mathcal{N}(0, \sigma^2)$. It has structural equation*

$$X_C^{\mathrm{do}(x_N)} = \alpha + \beta X_W^{\mathrm{do}(x_N)}.$$

For a given parameter θ, it has a solution function $\tilde{g}^\theta : (x_N, x_W) \mapsto \alpha + \beta x_W$. It has potential outcomes of the form $X_C^{\mathrm{do}(x_N)} = \alpha + \beta X_W^{\mathrm{do}(x_N)}$ for some random variable $X_W^{\mathrm{do}(x_N)} \sim \mathcal{N}(0, \sigma^2)$.

Remark 6.4.8. *We will interpret exogenous input variables in iSCMs always as if they represent a hard intervention with unspecified target values. While the target* ***values***

*are unspecified, this does imply certain assumptions regarding the data-generating **process** (that is, as to **how** the values can be chosen). Indeed, the data-generating process described in Section 6.3.6 implies that*

$$X_W^{(1)}, \ldots, X_W^{(n)} \perp\!\!\!\perp X_J^{(1)}, \ldots, X_J^{(n)}.$$

Here are two examples of settings in which this can be a realistic modeling assumption:

1. *If X_J is randomized (the agent that determines the value of $X_J^{(i)}$ just samples it from an independent source of randomness);*
2. *If the agent that determines the value of $X_J^{(i)}$ has no access to the values of $X_W^{(1)}, \ldots, X_W^{(n)}$ before or while deciding the value of $X_J^{(i)}$.*

In both cases, one needs to assume in addition that all generated samples end up in the data set (or weaker, that samples are observed 'completely at random': whether or not a sample is observed is a random variable that is independent of X_J and X_W).

Exercise 6.4.9. *Suppose you are interested in studying the relationship between age and body mass index. Is it a good idea to represent age as an exogenous input variable in an iSCM? Why, or why not?*

Summarizing, we have now seen three different ways of representing hard interventions, which are all 'hard' in the sense that they completely override the default causal mechanisms of their endogenous targets, so that their values are no longer determined by those of other endogenous variables. The three variants differ in how we decide to model the intervened variables: as exogenous inputs, as exogenous random variables, or as endogenous variables with a constant value.[38]

6.4.2. Soft interventions (mechanism changes)

Soft interventions (also known as mechanism changes) replace the causal mechanism of an endogenous variable by another causal mechanism.

Definition 6.4.10 (Soft intervention). *Given an iSCM $M = (J, V, W, \mathcal{X}, P, f)$, an intervention target $T \subseteq V$, and measurable functions $\tilde{f}_v : \mathcal{X} \to \mathcal{X}_v$ for $v \in T$, a **soft intervention on** M **replacing** f_T **by** $\tilde{f}_T$ yields an intervened iSCM of the form*

$$M_{\mathrm{do}(T \leftarrow \tilde{f}_T)} := \Big(J, V, W, \mathcal{X}, P, (f_{V \setminus T}, \tilde{f}_T)\Big)\,.$$

Proposition 6.4.11. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. (Hard or soft) interventions $\mathrm{do}(T_1 \ldots)$, $\mathrm{do}(T_2 \ldots)$ with disjoint targets $T_1, T_2 \subseteq J \cup V$ (of any of the four variants) commute:*

$$\big(M_{\mathrm{do}(T_1 \ldots)}\big)_{\mathrm{do}(T_2 \ldots)} = \big(M_{\mathrm{do}(T_2 \ldots)}\big)_{\mathrm{do}(T_1 \ldots)} = M_{\mathrm{do}((T_1 \cup T_2) \ldots)}.$$

Proof. This follows by writing out the definitions and checking commutativity of the operations performed on the various components of the iSCM tuple one-by-one. □

[38] Since most accounts on iSCMs do not provide for the possibility of exogenous input variables, the two variants most often seen in the literature are the latter two.

6.4.3. Intervention variables

It is often convenient to combine an iSCM and one or more intervened versions of the iSCM together in a single iSCM. This can be done by introducing *intervention variables* that indicate whether, and possibly encode how, an intervention is performed.

Example 6.4.12. *Consider an iSCM with endogenous variables X_1, X_2, exogenous random variables W_1, W_2 and structural equations:*

$$X_1 = f_1(X_2, W_1), \qquad X_2 = f_2(X_1, W_2).$$

After a hard intervention $\mathrm{do}(X_2 = x_2)$ the structural equations become:

$$X_1 = f_1(X_2, W_1), \qquad X_2 = x_2.$$

We can combine both into a single iSCM by adding an intervention variable I_2 (an exogenous input variable) and change the structural equations into:

$$X_1 = f_1(X_2, W_1), \qquad X_2 = \begin{cases} f_2(X_1, W_2) & I_2 = 0 \\ x_2 & I_2 = 1. \end{cases}$$

While there are many ways of intervening, and also many ways of encoding specific interventions, a special case that we will encounter frequently is the following way of using intervention variables to model hard interventions on subsets of B for some subset $B \subseteq V$ of endogenous variables.

Definition 6.4.13. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. Let $B \subseteq V \cup J$. For each $b \in B \cap V$, we introduce an additional 'intervention variable' I_b, which will become an exogenous input variable; for $j \in B \cap J$, we just set $I_j := j$. We denote $I_B := (I_b)_{b \in B}$. We define an extended iSCM*

$$M_{\mathrm{do}(I_B)} := \left(J \cup I_B, V, W, \mathcal{X} \times \prod_{b \in B \cap V} \mathcal{X}_{I_b}, P, \tilde{f} \right)$$

with $\mathcal{X}_{I_b} := \mathcal{X}_b \,\dot\cup\, \{\star\}$, with causal mechanism $\tilde{f}$ with components

$$\begin{aligned} v \in V \cap B: \qquad & \tilde{f}_v(x_{V \cup J \cup W \cup I_B}) := \begin{cases} f_v(x_{V \cup J \cup W}) & x_{I_v} = \star \\ x_{I_v} & x_{I_v} \in \mathcal{X}_v \end{cases} \\ v \in V \setminus B: \qquad & \tilde{f}_v(x_{V \cup J \cup W \cup I_B}) := f_v(x_{V \cup J \cup W}) \end{aligned}$$

For $b \in B \cap V$, $x_{I_b} = \star$ encodes that there is no intervention on b, while $x_{I_b} \neq \star$ encodes that the hard intervention $\mathrm{do}(X_b = x_{I_b})$ is performed.

This provides a way to simultaneously encode $\{M_{\mathrm{do}(X_C)} : C \subseteq B\}$ into a single iSCM $M_{\mathrm{do}(I_B)}$ and we will make use of it to derive the do-calculus (see Section 7.3).

We have the following commutation relations.

Proposition 6.4.14. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. Adding intervention variables to $b_1 \subseteq V \cup J$ and to $b_2 \subseteq V \cup J$ (with $b_1 \neq b_2$) commutes:*

$$(M_{\mathrm{do}(I_{b_1})})_{\mathrm{do}(I_{b_2})} = (M_{\mathrm{do}(I_{b_2})})_{\mathrm{do}(I_{b_1})} = M_{\mathrm{do}(I_{\{b_1,b_2\}})}.$$

Proof. This follows by writing out the definitions and checking commutativity of the operations performed on the various components of the iSCM tuple one-by-one. □

Proposition 6.4.15. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. Adding intervention variables to $B \subseteq V \cup J$ commutes with (hard or soft) interventions $\mathrm{do}(T \dots)$ with disjoint target $T \subseteq J \cup V$ (of any of the four variants):*

$$(M_{\mathrm{do}(I_B)})_{\mathrm{do}(T\dots)} = (M_{\mathrm{do}(T\dots)})_{\mathrm{do}(I_B)}.$$

Proof. This follows by writing out the definitions and checking commutativity of the operations performed on the various components of the iSCM tuple one-by-one. □

6.5. Composition and decomposition

If we think about an iSCM as modeling a 'system', then we also obtain a model for any 'subsystem' in the following way.

Definition 6.5.1 (Taking a submodel of an iSCM)**.** *Given an iSCM $M = (J, V, W, \mathcal{X}, P, f)$ and a subset $V' \subseteq V$ of its endogenous variables, we define its **submodel on** V' as the iSCM*

$$M^{[V']} := (J \cup (V \setminus V'), V', W, \mathcal{X}, P, f_{V'}) .$$

Remark 6.5.2. *Note that this is just the intervened iSCM $M_{\mathrm{do}(V \setminus V')}$.*

Given two iSCMs, such that some of the variables of one iSCM can be used as (part of the) exogenous input of the other, and possibly vice versa, we can compose them into a single iSCM.[39]

Definition 6.5.3 (Composing two iSCMs)**.** *Given two iSCMs $M = (J, V, W, \mathcal{X}, P, f)$ and $\tilde{M} = \left(\tilde{J}, \tilde{V}, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f}\right)$ and two subsets $C \subseteq J \cap (\tilde{V} \cup \tilde{W})$ and $\tilde{C} \subseteq (V \cup W) \cap \tilde{J}$ satisfying the following conditions:*

1. *for all $c \in C$, $(\mathcal{X})_c = (\tilde{\mathcal{X}})_c$,*
2. *for all $c \in \tilde{C}$, $(\mathcal{X})_c = (\tilde{\mathcal{X}})_c$,*
3. *$(J \setminus C) \cap (\tilde{J} \setminus \tilde{C}) = \emptyset$,*
4. *$V \cap \tilde{V} = \emptyset$,*

[39] In practice, one may have to relabel the variables first before one can perform this composition operation, by changing the index sets of the iSCMs to ensure that the right variables become coupled.

5. $W \cap \tilde{W} = \emptyset$,

we define the ***composed iSCM***

$$\mathrm{comp}(M, C, \tilde{M}, \tilde{C}) := \Big((J \setminus C) \,\dot\cup\, (\tilde{J} \setminus \tilde{C}), V \,\dot\cup\, \tilde{V}, W \,\dot\cup\, \tilde{W}, \mathcal{X}^\circ, P^\circ, f^\circ\Big),$$

where

$$\begin{aligned} \mathcal{X}^\circ &:= \mathcal{X}_{(J\setminus C)\,\dot\cup\, V \,\dot\cup\, W} \times (\tilde{\mathcal{X}})_{(\tilde{J}\setminus\tilde{C})\,\dot\cup\,\tilde{V}\,\dot\cup\,\tilde{W}}, \\ P^\circ &:= P \otimes \tilde{P}, \\ f^\circ &:= (f, \tilde{f}). \end{aligned}$$

The special case $C = \emptyset$ will also be denoted as $\tilde{M} \circ M$, and likewise the special case $\tilde{C} = \emptyset$ will be denoted as $M \circ \tilde{M}$.

One can consider the decomposition (taking submodels) as 'dual to' the composition (combining submodels). These operations formalize the notion of *modularity*, that is, an iSCM can be thought of modeling a system consisting of interacting components by modeling for each component separately how it interacts with the other components. The submodels on individual variables correspond with 'atomic' subsystems that cannot be (or won't be) decomposed into smaller parts.

6.6. Essentially Unique Solvability and Simple iSCMs

Whereas the existence of solution functions is important, their uniqueness also plays a pivotal role in the theory of (cyclic) iSCMs. A technical subtlety is that like existence, we will only demand uniqueness up to certain null sets.

Definition 6.6.1. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. We say that M is* ***essentially uniquely solvable*** *if there exists a measurable function $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ that satisfies*

$$x_V = f(x) \iff x_V = g(x_J, x_W) \quad M\text{-a.s.}.$$

For a subset $L \subseteq V$, we say that M is ***essentially uniquely solvable w.r.t.*** *L if there exists a measurable function $g^{[L]} : \mathcal{X}_J \times \mathcal{X}_{(V\setminus L)} \times \mathcal{X}_W \to \mathcal{X}_L$ that satisfies*

$$x_L = f_L(x) \iff x_L = g^{[L]}(x_J, x_{(V\setminus L)}, x_W) \quad M\text{-a.s.}.$$

Note that such a function g is a solution function of M, and $g^{[L]}$ is a partial solution function of M w.r.t. L. Hence, if M is essentially uniquely solvable (w.r.t. L), then M is solvable (w.r.t. L). In addition to the existence of (partial) uniqueness functions, their uniqueness up to certain null sets follows.

Definition 6.6.2. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. We say that its solution function is* ***essentially unique*** *if $g = \tilde{g}$ M-a.s. for any two solution functions $g, \tilde{g}$ of M. We say that its partial solution function w.r.t. $L \subseteq V$ is* ***essentially unique*** *if $g^{[L]} = \tilde{g}^{[L]}$ M-a.s. for any two partial solution functions $g^{[L]}, \tilde{g}^{[L]}$ of M w.r.t. L.*

The following observation is a simple consequence of the definitions that we will employ often.

Proposition 6.6.3. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. Suppose that M is essentially uniquely solvable w.r.t. $L \subseteq V$. Then for any (partial) solution function $\tilde{g}^{[L]}$ of M w.r.t. L we have:*

$$x_L = f_L(x) \iff x_L = \tilde{g}^{[L]}(x_J, x_{(V \setminus L)}, x_W) \quad M\text{-}a.s..$$

Hence, the (partial) solution function of M w.r.t. L is essentially unique.

Proof. Since M is essentially uniquely solvable w.r.t. L, there exists a measurable function $g^{[L]} : \mathcal{X}_J \times \mathcal{X}_{(V \setminus L)} \times \mathcal{X}_W \to \mathcal{X}_L$ that satisfies

$$x_L = f_L(x) \iff x_L = g^{[L]}(x_J, x_{(V \setminus L)}, x_W) \quad M\text{-a.s..}$$

Then

$$[\forall x_V \in \mathcal{X}_V : (x_L = f_L(x_J, x_{(V \setminus L)}, x_W, x_L) \iff x_L = g^{[L]}(x_J, x_{(V \setminus L)}, x_W))] \quad M\text{-a.s..}$$

Let $\tilde{g}^{[L]} : \mathcal{X}_J \times \mathcal{X}_{(V \setminus L)} \times \mathcal{X}_W \to \mathcal{X}_L$ be a partial solution function of M w.r.t. L. Then:

$$\tilde{g}^{[L]}(x_J, x_{(V \setminus L)}, x_W) = f_L(x_J, x_{(V \setminus L)}, x_W, \tilde{g}^{[L]}(x_J, x_{(V \setminus L)}, x_W)) \quad M\text{-a.s..}$$

Combining these we conclude that

$$g^{[L]}(x_J, x_{(V \setminus L)}, x_W) = \tilde{g}^{[L]}(x_J, x_{(V \setminus L)}, x_W) \quad M\text{-a.s..}$$

Together with the property of $g^{[L]}$, this also implies:

$$x_L = f_L(x) \iff x_L = \tilde{g}^{[L]}(x_J, x_{(V \setminus L)}, x_W) \quad M\text{-a.s..}$$

$\square$

Essentially uniquely solvable iSCMs not only have essentially unique solution functions, but also unique distributions of potential outcomes and unique Markov kernels of solutions.

Theorem 6.6.4. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. If M is essentially uniquely solvable, then:*

1. *Its solution function is essentially unique;*
2. *The distribution of a potential outcome $X_{V \cup W}^{\mathrm{do}(x_J)}$ for input $x_J \in X_J$ is unique and given by*
$$X_{V \cup W}^{\mathrm{do}(x_J)} \sim (g^{\mathrm{do}(x_J)}, \mathrm{id}_{\mathcal{X}_W})_* P_M(X_W)$$
where $g^{\mathrm{do}(x_J)} : \mathcal{X}_W \to \mathcal{X}_V : x_W \mapsto g(x_J, x_W)$ with $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ a solution function of M.

3. M has a unique Markov kernel given by

$$P_M(X_V, X_W \mid \mathrm{do}(X_J)) = (g, \mathrm{id}_{\mathcal{X}_W})_* \left(\left(\bigotimes_{w \in W} P_w(X_w) \right) \otimes \delta(X_J | X_J) \right)$$

for $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ a solution function of M.

Proof. If M is essentially uniquely solvable, it clearly has a solution function. By Proposition 6.3.17, M then also has potential outcomes and solutions. We proceed to show their (essential) uniqueness.

1. Let $x_J \in \mathcal{X}_J$ and $X_{V,W}^{\mathrm{do}(x_J)}$ a potential outcome of M with input x_J. Let g be a solution function of M. Then:

$$[x_V = f(x_J, x_W, x_V) \iff x_V = g(x_J, x_W)] \quad M\text{-a.s.}.$$

   With Lemma 6.3.13 we conclude:

$$[X_V^{\mathrm{do}(x_J)} = f(x_J, X_W^{\mathrm{do}(x_J)}, X_V^{\mathrm{do}(x_J)}) \iff X_V^{\mathrm{do}(x_J)} = g(x_J, X_W^{\mathrm{do}(x_J)})] \quad \text{a.s.}$$

   This implies:

$$[X_V^{\mathrm{do}(x_J)} = f(x_J, X_W^{\mathrm{do}(x_J)}, X^{\mathrm{do}(x_J)}) \text{ a.s.}] \iff [X_V^{\mathrm{do}(x_J)} = g(x_J, X_W^{\mathrm{do}(x_J)}) \text{ a.s.}].$$

   Since the l.h.s. holds by assumption, we conclude that

$$X_{V \cup W}^{\mathrm{do}(x_J)} = (g(x_J, X_W^{\mathrm{do}(x_J)}), X_W^{\mathrm{do}(x_J)}) \text{ a.s.},$$

   which in turn implies

$$X_{V \cup W}^{\mathrm{do}(x_J)} \sim (g^{\mathrm{do}(x_J)}, \mathrm{id}_{\mathcal{X}_W})_* P(X_W^{\mathrm{do}(x_J)})$$

   and therefore

$$X_{V \cup W}^{\mathrm{do}(x_J)} \sim (g^{\mathrm{do}(x_J)}, \mathrm{id}_{\mathcal{X}_W})_* P_M(X_W).$$

2. Let $K(X_V, X_W \mid X_J)$ denote the Markov kernel for M corresponding to a solution $X : \mathcal{U} \times \mathcal{X}_J \to \mathcal{X}$. For any $x_J \in \mathcal{X}_J$, the random variable $X^{\mathrm{do}(x_J)} : \mathcal{U} \to \mathcal{X} : u \mapsto X(u, x_J)$ is a potential outcome of M. Its distribution coincides with that of $K(X_V, X_W \mid X_J = x_J)$, by definition. The claim now follows from 3.

□

For the special case of no exogenous input variables ($J = \emptyset$), this means that essentially uniquely solvable SCMs induce a unique distribution.

If an iSCM is essentially uniquely solvable, this does not necessarily mean that it is still essentially uniquely solvable after performing some intervention. To avoid the complications introduced in case solutions are absent, or present but not unique, we will henceforth make strong assumptions regarding the existence and uniqueness of solutions. We will (mostly) restrict our attention to a subclass of iSCMs that we refer to as *simple iSCMs*, which are iSCMs that are essentially uniquely solvable and remain so after any hard intervention:

Definition 6.6.5. *An iSCM $M = (J, V, W, \mathcal{X}, P, f)$ is called* ***simple*** *if for all $T \subseteq V$, M is essentially uniquely solvable w.r.t. T.*

Note that this includes essentially unique solvability of M itself for $T = \emptyset$.

Remark 6.6.6. *An equivalent condition for M to be essentially uniquely solvable w.r.t. L is provided by Proposition 6.3.19. Thereby, M is a simple iSCM if and only if for each $L \subseteq V$ the following holds:*

$$[\exists! x_L \in \mathcal{X}_L : x_L = f_L(x_J, x_{V \setminus L}, x_W, x_L)]$$

up to a measurable M-null set.

For M an SCM without input nodes (that is, $J = \emptyset$), every M-null set is contained in a measurable M-null set if $J = \emptyset$. Hence, in that case Proposition 6.3.19 implies that M is essentially uniquely solvable if and only if

$$\exists! x_V \in \mathcal{X}_V : x_V = f(x_V, x_W) \quad P\text{-a.s.}.$$

Example 6.6.7. *Consider an iSCM with structural equations*

$$\begin{aligned}
X_1 &= W_1 \\
X_2 &= W_2 \\
X_3 &= X_1 X_4 + W_3 \\
X_4 &= X_2 X_3 + W_4
\end{aligned}$$

where the X's are considered real-valued endogenous variables and the W's exogenous variables with domains $(-1, 1) \subset \mathbb{R}$. We can solve the system of structural equations for X in terms of W:

$$\begin{aligned}
X_1 &= W_1 \\
X_2 &= W_2 \\
X_3 &= \frac{W_3 + W_1 W_4}{1 - W_1 W_2} \\
X_4 &= \frac{W_2 W_3 + W_4}{1 - W_2 W_1}
\end{aligned}$$

which gives a unique solution function. Similarly, we can take any subset of the structural equations and solve it for the variables appearing on the l.h.s. of the equations in the subset, and obtain a unique solution function. For example, only solving the structural equations for X_3 and X_4, we obtain:

$$\begin{aligned}
X_3 &= \frac{W_3 + W_1 W_4}{1 - W_1 X_2} \\
X_4 &= \frac{W_2 W_3 + W_4}{1 - W_2 X_1}
\end{aligned}$$

where the variables X_1 and X_2 are now considered as exogenous input variables (instead of endogenous variables). Hence, any subset of the structural equations has a unique solution for the variables appearing on the l.h.s. in terms of the remaining ones on the r.h.s., which means that this iSCM is simple.

If we extend the range of possible values of the exogenous random variables to $\mathcal{X}_W = \mathbb{R}^2$, then we may still end up with essentially unique solution functions as long as $\{(w_1, w_2) \in \mathbb{R}^2 : 1 = w_1 w_2\}$ is a measurable M-null set (which depends on the distribution $P(X_W)$ we would pick).

One of the very convenient properties of simple iSCMs is that after any hard intervention, they still have a uniquely defined Markov kernel. Simplicity is preserved by hard interventions and by adding intervention variables.

Proposition 6.6.8. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Then:*

1. *$M_{\mathrm{do}(T\ldots)}$ is simple for $T \subseteq V \cup J$ (hard interventions of all three types);*
2. *$M_{\mathrm{do}(I_B)}$ is simple for $B \subseteq V \cup J$ (adding intervention variables).*

Proof. Because M is simple, for all $T \subseteq V$ there exists a measurable function $g_{\mathrm{do}(T)} : \mathcal{X}_{J \cup T} \times \mathcal{X}_W \to \mathcal{X}_{V \setminus T}$ such that:

$$x_{V \setminus T} = f_{V \setminus T}(x) \iff x_{V \setminus T} = g_{\mathrm{do}(T)}(x_J, x_T, x_W) \quad M\text{-a.s.}.$$

Let $\tilde{M} = \langle \tilde{J}, \tilde{V}, \tilde{W}, \mathcal{X}, \tilde{P}, \tilde{f} \rangle$ denote an intervened iSCM (we will consider special cases below). To prove that $\tilde{M}$ is simple, it suffices to show that for all $\tilde{T} \subseteq \tilde{V}$ there exists a measurable function $\tilde{g}_{\mathrm{do}(\tilde{T})} : \mathcal{X}_{\tilde{J} \cup \tilde{T}} \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_{\tilde{V} \setminus \tilde{T}}$ such that:

$$x_{\tilde{V} \setminus \tilde{T}} = \tilde{f}_{\tilde{V} \setminus \tilde{T}}(x) \iff x_{\tilde{V} \setminus \tilde{T}} = \tilde{g}_{\mathrm{do}(\tilde{T})}(x_{\tilde{J}}, x_{\tilde{T}}, x_{\tilde{W}}) \quad \tilde{M}\text{-a.s.},$$

as long as M-a.s. implies $\tilde{M}$-a.s..

1. (i) Let $T \subseteq J \cup V$. Consider $\tilde{M} = M_{\mathrm{do}(T)}$, that is, $\tilde{J} = J \cup T$, $\tilde{V} = V \setminus T$, $\tilde{W} = W$, $\tilde{P} = P$, $\tilde{f} = f_{\tilde{V}}$. Since $\tilde{M}_{\mathrm{do}(\tilde{T})} = M_{\mathrm{do}(T \cup \tilde{T})}$ for $\tilde{T} \subseteq \tilde{V}$, we can just take $\tilde{g}_{\mathrm{do}(\tilde{T})} = g_{\mathrm{do}((T \setminus J) \cup \tilde{T})}$. Note that indeed M-a.s. implies $M_{\mathrm{do}(T)}$-a.s.. Hence $M_{\mathrm{do}(T)}$ is simple.

   (ii) Let $T \subseteq V \cup J$, and $\xi_T \in \mathcal{X}_T$. Consider $\tilde{M} = M_{\mathrm{do}(x_T = \xi_T)}$, that is, $\tilde{J} = J \setminus T$, $\tilde{V} = V \cup T$, $\tilde{W} = W$, $\tilde{P} = P$, $\tilde{f} = (f_{V \setminus T}, \xi_T)$. Since for $\tilde{T} \subseteq \tilde{V}$:

   $$\tilde{M}_{\mathrm{do}(\tilde{T})} = (M_{\mathrm{do}(x_{T \setminus \tilde{T}} = \xi_{T \setminus \tilde{T}})})_{\mathrm{do}(\tilde{T})},$$

   as solution function $\tilde{g}_{\mathrm{do}(\tilde{T})}$ we can take the one for $M_{\mathrm{do}(T \cup \tilde{T})}$, evaluated in $x_{T \setminus \tilde{T}} = \xi_{T \setminus \tilde{T}}$, and with additional output $T \setminus \tilde{T}$ with value $\xi_{T \setminus \tilde{T}}$:

   $$\tilde{g}_{\mathrm{do}(\tilde{T})}(x_{\tilde{J}}, x_{\tilde{T}}, x_W) = \big(g_{\mathrm{do}((T \setminus J) \cup \tilde{T})}(\xi_{T \setminus \tilde{T}}, x_{\tilde{T} \cup J}, x_W), \xi_{T \setminus \tilde{T}}\big).$$

   Note that indeed M-a.s. implies $M_{\mathrm{do}(x_T = \xi_T)}$-a.s.. Hence $M_{\mathrm{do}(x_T = \xi_T)}$ is simple.

(iii) Let $T \subseteq V \cup J$ and $Q_T \in \bigotimes_{t \in T} \mathcal{P}(\mathcal{X}_t)$. Consider $\tilde{M} = M_{\mathrm{do}(X_T \sim Q_T)}$, that is, $\tilde{J} = J \setminus T$, $\tilde{V} = V \setminus T$, $\tilde{W} = W \cup T$, $\tilde{P} = P \otimes Q_T$, $\tilde{f} = f_{V \setminus T}$. Since for $\tilde{T} \subseteq \tilde{V}$:

$$\tilde{M}_{\mathrm{do}(\tilde{T})} = (M_{\mathrm{do}(T \cup \tilde{T})})_{\mathrm{do}(X_T \sim Q_T)}$$

we can take $\tilde{g}_{\mathrm{do}(\tilde{T})} = g_{\mathrm{do}((T \setminus J) \cup \tilde{T})}$. Note that indeed M-a.s. implies $M_{\mathrm{do}(X_T \sim Q_T)}$-a.s.. Hence $M_{\mathrm{do}(X_T \sim Q_T)}$ is simple.

2. For $B \subseteq J$, the operation of adding intervention variables $\mathrm{do}(I_B)$ only adds indices $I_j := j$ for $j \in B \cap J$. Therefore, we can consider $B \subseteq V$ without loss of generality. Furthermore, by using induction we only need to consider the case $B = \{b\} \subseteq V$ by means of Proposition 6.4.14. Hence consider $\tilde{M} = M_{\mathrm{do}(I_b)}$, that is, $\tilde{J} = J \cup \{I_b\}$, $\tilde{V} = V$, $\tilde{W} = W$, $\tilde{\mathcal{X}} = \mathcal{X} \times \mathcal{X}_{I_b}$ with $\mathcal{X}_{I_b} := \mathcal{X}_b \dot{\cup} \{\star\}$, $\tilde{P} = P$, and $\tilde{f}$ given by

$$\begin{aligned} \tilde{f}_b(x_{I_b}, x_{V \cup J \cup W}) &:= \begin{cases} f_b(x_{V \cup J \cup W}) & x_{I_b} = \star \\ x_{I_b} & x_{I_b} \in \mathcal{X}_b \end{cases} \\ v \in V \setminus \{b\}: \qquad \tilde{f}_v(x_{I_b}, x_{V \cup J \cup W}) &:= f_v(x_{V \cup J \cup W}) \end{aligned}$$

Let $\tilde{T} \subseteq \tilde{V}$. Consider first the case $b \notin \tilde{T}$. Then:

$$\tilde{M}_{\mathrm{do}(\tilde{T})} = (M_{\mathrm{do}(\tilde{T})})_{\mathrm{do}(I_b)}.$$

We can take:

$$\tilde{g}_{\mathrm{do}(\tilde{T})}(x_{I_b}, x_J, x_{\tilde{T}}, x_W) = \begin{cases} g_{\mathrm{do}(\tilde{T})}(x_J, x_{\tilde{T}}, x_W) & x_{I_b} = \star \\ (g_{\mathrm{do}(\tilde{T} \cup \{b\})}(x_J, x_{\tilde{T}}, x_b = x_{I_b}, x_W), x_{I_b}) & x_{I_b} \in \mathcal{X}_b \end{cases}$$

as one can check that

$$\begin{cases} x_{V \setminus \tilde{T}} = f_{V \setminus \tilde{T}}(x_{J \cup W \cup V}) \iff x_{V \setminus \tilde{T}} = g_{\mathrm{do}(\tilde{T})}(x_J, x_{\tilde{T}}, x_W) \quad M\text{-a.s.} . \\ x_{V \setminus (\tilde{T} \cup \{b\})} = f_{V \setminus (\tilde{T} \cup \{b\})}(x_{J \cup W \cup V}) \iff x_{V \setminus (\tilde{T} \cup \{b\})} = g_{\mathrm{do}(\tilde{T} \cup \{b\})}(x_J, x_{\tilde{T} \cup \{b\}}, x_W) \quad M\text{-a.s.} . \end{cases}$$

implies that for each $x_{I_b} \in \mathcal{X}_{I_b}$:

$$x_{V \setminus \tilde{T}} = \tilde{f}_{V \setminus \tilde{T}}(x_{I_b}, x_{J \cup W \cup V}) \iff x_{V \setminus \tilde{T}} = \tilde{g}_{\mathrm{do}(\tilde{T})}(x_{I_b}, x_J, x_{\tilde{T}}, x_W) \quad M\text{-a.s.} .$$

Now consider the case $b \in \tilde{T}$. Then:

$$\tilde{M}_{\mathrm{do}(\tilde{T})} = ((M_{\mathrm{do}(\tilde{T} \setminus \{b\})})_{\mathrm{do}(I_b)})_{\mathrm{do}(b)}.$$

This is just $M_{\mathrm{do}(\tilde{T})}$ with an additional 'unused' exogenous input variable I_b. We can therefore take:

$$\tilde{g}_{\mathrm{do}(\tilde{T})}(x_{I_b}, x_J, x_{\tilde{T}}, x_W) = g_{\mathrm{do}(\tilde{T})}(x_J, x_{\tilde{T}}, x_W)$$

where one should notice that in this case the desired property holds $M_{\mathrm{do}(I_b)}$-a.s., as it does not depend on x_{I_b} and holds M-a.s..

This completes the proof. □

Simplicity is in general not preserved by soft interventions.

Corollary 6.6.9. *If an iSCM $M = (J, V, W, \mathcal{X}, P, f)$ is simple, then its Markov kernel $P_M(X_V, X_W \mid \mathrm{do}(X_J))$ and all intervened Markov kernels*

$$P_{M_{\mathrm{do}(T)}}(X_{V\setminus T}, X_W \mid \mathrm{do}(X_{J\cup T}))$$

for $T \subseteq V$ exist and are unique.

Proof. If M is simple, then for all $T \subseteq V$ also $M_{\mathrm{do}(T)}$ is simple (by Proposition 6.6.8) and hence essentially uniquely solvable. The claim now follows from Theorem 6.6.4. □

We will make use of the following notation for intervened Markov kernels of simple iSCMs.

Notation 6.6.10. *For a simple iSCM M, we write for $T \subseteq V$:*

$$P_M(X_{V\setminus T}, X_W \mid \mathrm{do}(X_J), \mathrm{do}(X_T)) := P_{M_{\mathrm{do}(T)}}(X_{V\setminus T}, X_W \mid \mathrm{do}(X_{J\cup T}))$$

We establish the following terminology for these Markov kernels:

special case	*Markov kernel*	*name*
	$P_M(X_{V\setminus T} \mid \mathrm{do}(X_J), \mathrm{do}(X_T))$	***interventional Markov kernels of*** M
$T = \emptyset$	$P_M(X_V \mid \mathrm{do}(X_J))$	***observable Markov kernel of*** M
$J = T = \emptyset$	$P_M(X_V)$	***observational Markov kernel of*** M

The fact that these iSCMs are relatively simple to deal with (because we do not have to worry about non-existence or non-uniqueness of solutions) motivated their name. Even though simplicity is a strong assumption, the class of simple iSCMs is more expressive than the class of L-CBNs, since it allows to model causal cycles to some extent. In paricular, some of the examples at the beginning of this Chapter (in Section 6.1) can be dealt with using the class of simple iSCMs.

Exercise 6.6.11. *Show that Example 6.1.1 provides no difficulties when modeling using simple iSCMs. Is the class of iSCMs generally closed under merging variables?*

Exercise 6.6.12. *Show that the mixture model in Example 6.1.2 can be modeled as a simple iSCM. Is the class of iSCMs generally closed under taking mixtures of iSCMs?*

6.6.1. Linear iSCMs

For iSCMs whose causal mechanism is linear in terms of the endogenous variables (up to a certain null set), we can give a sufficient condition for essentially unique solvability, and explicitly write down the form of their solution function.

Definition 6.6.13. *Let $\mathbb{K}$ be a field (e.g., $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$). Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, where each $\mathcal{X}_v$ is a finite-dimensional Polish vector space over $\mathbb{K}$, and let $\mathcal{X}_A = \bigoplus_{v \in A} \mathcal{X}_a$ for $A \subseteq V$. Then M is called* ***essentially linear*** *if each component of the causal mechanism is (up to an M-null set) an affine combination of endogenous variables with coefficients that may be functions of the exogenous variables, i.e., of f is of the form*

$$f(x_J, x_W, x_V) = B(x_J, x_W)x_V + c(x_J, x_W) \qquad M\text{-a.s.},$$

where for all $x_J \in \mathcal{X}_J$ and all $x_W \in \mathcal{X}_W$, $B(x_J, x_W) : \mathcal{X}_V \to \mathcal{X}_V$ is linear and $c(x_J, x_W) \in \mathcal{X}_V$. Since $\mathcal{X}_V = \bigoplus_{v \in V} \mathcal{X}_v$, we can also write this as

$$f_v(x_J, x_W, x_V) = \sum_{u \in V} B_{vu}(x_J, x_W)x_u + c_v(x_J, x_W) \qquad M\text{-a.s.},$$

where for all $x_J \in \mathcal{X}_J$ and all $x_W \in \mathcal{X}_W$, $B_{vu}(x_J, x_W) : \mathcal{X}_u \to \mathcal{X}_v$ is linear and $c_v(x_J, x_W) \in \mathcal{X}_v$ for all $u, v \in V$.

For $\mathbb{K} = \mathbb{R}$ and $\mathcal{X}_v = \mathbb{R}$ for all $v \in V$, the mapping B can be thought of as a matrix-valued measurable function $\mathcal{X}_J \times \mathcal{X}_W \to \mathbb{R}^{V \times V}$ and c as a vector-valued function $\mathcal{X}_J \times \mathcal{X}_W \to \mathbb{R}^V$.

For essentially linear iSCMs, essentially unique solvability is implied by the invertability of a certain matrix.

Proposition 6.6.14. *Let M be a essentially linear iSCM as in Definition 6.6.13. Denote by $I \in \mathbb{K}^{V \times V}$ the identity matrix. Then, M is essentially uniquely solvable if the matrices $I - B(x_J, x_W)$ are invertible for all $x_J \in \mathcal{X}_J, x_W \in \mathcal{X}_W$ up to a measurable M-null set. The corresponding essentially unique solution function is:*

$$g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V : (x_J, x_W) \mapsto (I - B(x_J, x_W))^{-1} c(x_J, x_W).$$

More generally, let $L \subseteq V$. Denote by $I_L \in \mathbb{R}^{L \times L}$ the identity matrix and $B_{LL}(x_J, x_W) \in \mathbb{K}^{L \times L}$ the submatrix of $B(x_J, x_W)$. M is essentially uniquely solvable w.r.t. L if the matrices $I_L - B_{LL}(x_J, x_W)$ are invertible for all $x_J \in \mathcal{X}_J, x_W \in \mathcal{X}_W$ up to a measurable M-null set. The corresponding essentially unique partial solution function w.r.t. L is:

$$\begin{aligned} g^{[L]} &: \mathcal{X}_J \times \mathcal{X}_{V \setminus L} \times \mathcal{X}_W \to \mathcal{X}_L \\ &: (x_J, x_{V \setminus L}, x_W) \mapsto (I_L - B_{LL}(x_J, x_W))^{-1} \big(B_{L, V \setminus L}(x_J, x_W) x_{V \setminus L} + c_L(x_J, x_W)\big). \end{aligned}$$

6.7. Equivalence Notions

In this section, we will discuss several important equivalence relations between iSCMs.

M and $\tilde{M}$ are considered equivalent if they differ only in terms of their causal mechanisms, yet each of their structural equations has almost surely the same solutions.

Definition 6.7.1. *Let $M = (J, V, W, \mathcal{X}, P, f)$ and $\tilde{M} = \left(J, V, W, \mathcal{X}, P, \tilde{f}\right)$ be two iSCMs that may differ only in terms of their causal mechanism. We say that M is* ***equivalent*** *to $\tilde{M}$ and write $M \equiv \tilde{M}$ if for each $v \in V$,*

$$x_v = f_v(x_J, x_V, x_W) \iff x_v = \tilde{f}_v(x_J, x_V, x_W) \qquad M\text{-a.s.}$$

(note that M-a.s. means the same as $\tilde{M}$-a.s.).

Equivalent iSCMs are indeed equivalent for many purposes, and it is compatible with all properties of and operations on iSCMS that we have defined thus far.

Proposition 6.7.2. *1. Equivalence is preserved by interventions: if $M \equiv \tilde{M}$ then:*

a) $M_{\mathrm{do}(T\dots)} \equiv \tilde{M}_{\mathrm{do}(T\dots)}$ for $T \subseteq V \cup J$ (hard interventions of all three types);

b) $M_{\mathrm{do}(T \leftarrow \hat{f}_T)} \equiv \tilde{M}_{\mathrm{do}(T \leftarrow \hat{f}_T)}$ for $T \subseteq V$ and $\hat{f}_T : \mathcal{X}_J \times \mathcal{X}_W \times \mathcal{X}_V \to \mathcal{X}_T$ a measurable function (soft interventions);

c) $M_{\mathrm{do}(I_B)} \equiv \tilde{M}_{\mathrm{do}(I_B)}$ for $B \subseteq V \cup J$ (adding intervention variables);

2. Equivalent iSCMs have equivalent submodels and composition preserves equivalence;

3. Equivalent iSCMs have the same null sets: if $M \equiv \tilde{M}$ then N is an M-null set if and only if it is an $\tilde{M}$-null set, and 'M-a.s.' means the same as '$\tilde{M}$-a.s.';

4. Equivalent iSCMs have the same potential outcomes, solutions, Markov kernels, and (partial) solution functions;

5. (Partial) solvability, essentially unique (partial) solvability, essential uniqueness of (partial) solution functions and simplicity are invariants under equivalence: if $M \equiv \tilde{M}$ then

a) M is solvable w.r.t. $L \subseteq V$ if and only if $\tilde{M}$ is solvable w.r.t. L,

b) M is essentially uniquely solvable w.r.t. $L \subseteq V$ if and only if it $\tilde{M}$ is essentially uniquely solvable w.r.t. L,

c) $g^{[L]}$ is an essentially unique solution function of M w.r.t. $L \subseteq V$ if and only if it is an essentially unique solution function of $\tilde{M}$ w.r.t. L,

d) M is simple if and only if $\tilde{M}$ is simple;

6. Essential linearity is preserved.

Proof. The first three claims follow directly from the definitions.

We prove the fourth point. $M \equiv \tilde{M}$ implies that

$$x_V = f(x) \iff x_V = \tilde{f}(x) \ M\text{-a.s..} \tag{41}$$

- Potential outcomes are invariant. Let $x_J \in \mathcal{X}_J$. We show that if a random variable $x_{V,W}^{\mathrm{do}(x_J)}$ is a potential outcome of M, then it is a potential outcome of $\tilde{M}$. Using Lemma 6.3.13, (41) implies:

$$X_V^{\mathrm{do}(x_J)} = f(x_J, X_W^{\mathrm{do}(x_J)}, X_V^{\mathrm{do}(x_J)}) \iff X_V^{\mathrm{do}(x_J)} = \tilde{f}(x_J, X_W^{\mathrm{do}(x_J)}, X_V^{\mathrm{do}(x_J)}) \text{ a.s..}$$

  Since $X_W^{\mathrm{do}(x_J)} \sim P_M(X_W)$, this proves the claim.

- Solutions are invariant. Let $(\mathcal{U} \times \mathcal{X}_J, K(U|X_J))$ be a transition probability space and $X : \mathcal{U} \times \mathcal{X}_J \to \mathcal{X}$ be a conditional random variable. We show that if X is a solution of M, then it is a solution of $\tilde{M}$. Using Lemma 6.3.14, (41) implies:

$$X_V = f(X_J, X_W, X_V) \iff X_V = \tilde{f}(X_J, X_W, X_V) \text{ a.s..}$$

  Since $K(X_W|X_J) = P_M(X_W)$, this proves the claim.

- Markov kernels are invariant: this follows from the invariance of solutions.

- Partial solution functions are invariant. Let $L \subseteq V$. We show that if $g^{[L]}$ is a partial solution function of M w.r.t. L, then it is a partial solution function of $\tilde{M}$ w.r.t L. By assumption, $g^{[L]} : \mathcal{X}_J \times \mathcal{X}_{V\setminus L} \times \mathcal{X}_W \to \mathcal{X}_L$ is a measurable function. Because M-a.s. and $\tilde{M}$-a.s. mean the same,

$$g^{[L]}(x_J, x_{V\setminus L}, x_W) = f_L\big(x_J, x_{V\setminus L}, g^{[L]}(x_J, x_{V\setminus L}, x_W), x_W\big) \quad M\text{-a.s.}$$

  together with (41) implies:

$$g^{[L]}(x_J, x_{V\setminus L}, x_W) = \tilde{f}_L\big(x_J, x_{V\setminus L}, g^{[L]}(x_J, x_{V\setminus L}, x_W), x_W\big) \quad \tilde{M}\text{-a.s.,}$$

  which proves the claim.

The fifth point follows straightforwardly because M and $\tilde{M}$ have the same partial solution functions if they are equivalent.

The preservation of essential linearity follows because the coefficients of an affine mapping are identified if the mapping is known up to an M-null set. □

Example 6.7.3. *An iSCM with the single structural equation*

$$X = -X^3 + X + W$$

where X is endogenous, and W is exogenous, is equivalent to the iSCM obtained by replacing the structural equation with

$$X = \sqrt[3]{W}.$$

Note that X no longer appears on the r.h.s..

We often make use of weaker notions of equivalence as well. For simplicity of exposition, we provide the definitions only for simple iSCMs (the general definitions are provided in [BFPM21] for SCMs).

Definition 6.7.4. *Let $M = (J, V, W, \mathcal{X}, P, f)$ and $\tilde{M} = (\tilde{J}, \tilde{V}, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ be two simple iSCMs and $O \subseteq V \cap \tilde{V}$ a subset. We say that:*

1. *M and $\tilde{M}$ are* ***observably equivalent w.r.t.*** *O if $\mathcal{X}_O = \tilde{\mathcal{X}}_O$, $\mathcal{X}_{J\cap\tilde{J}} = \tilde{\mathcal{X}}_{J\cap\tilde{J}}$ and their marginal Markov kernels coincide:*

$$P_M(X_O \mid \mathrm{do}(X_J)) = P_{\tilde{M}}(X_O \mid \mathrm{do}(X_{\tilde{J}})).$$

   This has to be interpreted as both Markov kernels being a version of a Markov kernel $\mathcal{X}_{J\cap\tilde{J}} \dashrightarrow \mathcal{X}_O$, i.e., $P_M(X_O \mid \mathrm{do}(X_J))$ must be constant in $x_{J\setminus\tilde{J}}$ and $P_{\tilde{M}}(X_O \mid \mathrm{do}(X_{\tilde{J}}))$ must be constant in $x_{\tilde{J}\setminus J}$.[40]

2. *M and $\tilde{M}$ are* ***interventionally equivalent w.r.t.*** *O if for every subset $T \subseteq O$ the intervened iSCMs $M_{\mathrm{do}(T)}$ and $\tilde{M}_{\mathrm{do}(T)}$ are observably equivalent w.r.t. $O \setminus T$.*

If $O = V \cap \tilde{V}$, we may omit the qualitifier "w.r.t. O". If $J = \tilde{J} = \emptyset$, we also use the term ***observationally equivalent w.r.t.*** *O interchangebly with* ***observably equivalent w.r.t.*** *O.*

Remark 6.7.5. *Note that M and $\tilde{M}$ are interventionally equivalent w.r.t. $O \subseteq (V \cap \tilde{V})$ if and only if $\mathcal{X}_O = \tilde{\mathcal{X}}_O$, $\mathcal{X}_{J\cap\tilde{J}} = \tilde{\mathcal{X}}_{J\cap\tilde{J}}$ and for all $T \subseteq O$:*

$$P_M(X_O \mid \mathrm{do}(X_J, X_T)) = P_{\tilde{M}}(X_O \mid \mathrm{do}(X_{\tilde{J}}, X_T)).$$

More generally, one could define interventional equivalence not only with respect to an observed set of variables, but also with respect to a given set of interventions.

Equivalence is stronger than interventional equivalence, which in turn is stronger than observable equivalence.

Proposition 6.7.6. *For simple iSCMs $M, \tilde{M}$ and a subset $O \subseteq V \cap \tilde{V}$:*

1. *If $M \equiv \tilde{M}$ then M and $\tilde{M}$ are interventionally equivalent w.r.t. O.*
2. *Interventional equivalence of M and $\tilde{M}$ w.r.t. O implies observable equivalence of M and $\tilde{M}$ w.r.t. O.*

Proof.
1. Suppose $M \equiv \tilde{M}$. Then for every $T \subseteq O$, $M_{\mathrm{do}(T)} \equiv \tilde{M}_{\mathrm{do}(T)}$. Since equivalent simple iSCMs have the same Markov kernels, $M_{\mathrm{do}(T)}$ and $\tilde{M}_{\mathrm{do}(T)}$ are observably equivalent w.r.t. $O \setminus T$ for every $T \subseteq O$.
2. This is trivial (consider the intervention targeting $T = \emptyset$).

□

However, the reverse implication does not hold in general. This expresses that causal modeling is more refined than probabilistic modeling.

[40] In other words,

$$P_M(X_O \mid \mathrm{do}(X_J)) = P_M(X_O \mid \mathrm{do}(X_{J\cap\tilde{J}}, \cancel{X_{J\setminus\tilde{J}}})) = P_{\tilde{M}}(X_O \mid \mathrm{do}(X_{J\cap\tilde{J}}, \cancel{X_{\tilde{J}\setminus J}})) = P_{\tilde{M}}(X_O \mid \mathrm{do}(X_{\tilde{J}})).$$

Example 6.7.7 (Observable equivalence does not imply interventional equivalence). *Consider the SCM M with*

$$\begin{aligned} W &\sim \mathcal{N}(\mu, \sigma^2), \\ N &= W, \\ C &= \alpha + \beta N, \end{aligned}$$

and the SCM $\tilde{M}$ with

$$\begin{aligned} W &\sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2), \\ C &= W, \\ N &= \tilde{\alpha} + \tilde{\beta} C. \end{aligned}$$

These SCMs are simple and their distributions are respectively

$$P_M(N, C) = \mathcal{N}\left(\begin{pmatrix} \mu \\ \alpha + \beta\mu \end{pmatrix}, \begin{pmatrix} \sigma^2 & \beta\sigma^2 \\ \beta\sigma^2 & \beta^2\sigma^2 \end{pmatrix} \right)$$

and

$$P_{\tilde{M}}(N, C) = \mathcal{N}\left(\begin{pmatrix} \tilde{\alpha} + \tilde{\beta}\tilde{\mu} \\ \tilde{\mu} \end{pmatrix}, \begin{pmatrix} \tilde{\beta}^2\tilde{\sigma}^2 & \tilde{\beta}\tilde{\sigma}^2 \\ \tilde{\beta}\tilde{\sigma}^2 & \tilde{\sigma}^2 \end{pmatrix} \right)$$

For certain parameter choices, they are observably equivalent (to be precise, they are observable equivalent iff $\mu = \tilde{\alpha} + \tilde{\beta}\tilde{\mu}$ and $\tilde{\mu} = \alpha + \beta\mu$ and $\sigma^2 = \tilde{\beta}^2\tilde{\sigma}^2$ and $\beta\sigma^2 = \tilde{\beta}\tilde{\sigma}^2$ and $\beta^2\sigma^2 = \tilde{\sigma}^2$). However, they are not interventionally equivalent except for very special parameter choices (to be precise, they are interventionally equivalent iff $\beta = \tilde{\beta} = 0$ and $\sigma^2 = \tilde{\sigma}^2 = 0$ and $\mu = \tilde{\alpha}$ and $\tilde{\mu} = \alpha$).

We have seen that equivalence is preserved under various interventions. This is also the case for interventional equivalence (although one has to be careful with respect to which subset).

Proposition 6.7.8. *Assume that M and $\tilde{M}$ are interventionally equivalent iSCMs w.r.t. $O \subseteq (V \cap \tilde{V})$. Then, for $T \subseteq O \cup (J \cap \tilde{J})$:*

1. *$M_{\mathrm{do}(T)}$ and $\tilde{M}_{\mathrm{do}(T)}$ are interventionally equivalent w.r.t. $O \setminus T$ (hard interventions with unspecified target value);*
2. *$M_{\mathrm{do}(X_T=\xi_T)}$ and $\tilde{M}_{\mathrm{do}(X_T=\xi_T)}$ are interventionally equivalent w.r.t. $O \cup T$, for any $\xi_T \in \mathcal{X}_T$ (hard intervention with specified target value);*
3. *$M_{\mathrm{do}(X_T \sim Q_T)}$ and $\tilde{M}_{\mathrm{do}(X_T \sim Q_T)}$ are interventionally equivalent w.r.t. $O \cup T$, for any $Q_T \in \mathcal{P}(\mathcal{X}_T)$ (hard intervention with stochastic target value);*
4. *$M_{\mathrm{do}(I_T)}$ and $\tilde{M}_{\mathrm{do}(I_T)}$ are interventionally equivalent w.r.t. O (adding intervention variables).*

Proof. 1. This follows from the commutativity of hard interventions (Proposition 6.4.11), plus some bookkeeping, as we show now in detail. Since M and $\tilde{M}$ are interventionally equivalent w.r.t. O, for all subsets $S \subseteq O$, $M_{\mathrm{do}(S)}$ and $\tilde{M}_{\mathrm{do}(S)}$ are observably equivalent w.r.t. $O \setminus S$. Let $T \subseteq O \cup (J \cap \tilde{J})$. $M_{\mathrm{do}(T)}$ and $\tilde{M}_{\mathrm{do}(T)}$ are interventionally equivalent w.r.t. $O \setminus T$ if for all $\tilde{T} \subseteq O \setminus T$, $(M_{\mathrm{do}(T)})_{\mathrm{do}(\tilde{T})}$ and $(\tilde{M}_{\mathrm{do}(T)})_{\mathrm{do}(\tilde{T})}$ are observably equivalent w.r.t. $(O \setminus T) \setminus \tilde{T}$. That is, if $M_{\mathrm{do}(T \cup \tilde{T})}$ and $\tilde{M}_{\mathrm{do}(T \cup \tilde{T})}$ are observably equivalent w.r.t. $O \setminus (T \cup \tilde{T})$ for all $\tilde{T} \subseteq O \setminus T$. That is, if $M_{\mathrm{do}((T \cup \tilde{T}) \setminus (J \cap \tilde{J}))}$ and $\tilde{M}_{\mathrm{do}((T \cup \tilde{T}) \setminus (J \cap \tilde{J}))}$ are observably equivalent w.r.t. $O \setminus ((T \cup \tilde{T}) \setminus (J \cap \tilde{J}))$ for all $\tilde{T} \subseteq O \setminus T$. Note that for all $\tilde{T} \subseteq O \setminus T$, $(T \cup \tilde{T}) \setminus (J \cap \tilde{J}) \subseteq O$.

For the next two cases, $\mathcal{X}_O = \tilde{\mathcal{X}}_O$ and $\mathcal{X}_{J \cap \tilde{J}} = \tilde{\mathcal{X}}_{J \cap \tilde{J}}$ holds since M and $\tilde{M}$ are interventionally equivalent w.r.t. O. This same property needs to hold for the intervened iSCMs $M_{\mathrm{do}(T\ldots)}$ and $\tilde{M}_{\mathrm{do}(T\ldots)}$.

2. Let $T \subseteq O \cup (J \cap \tilde{J})$ and $\xi_T \in \mathcal{X}_T$. To show: $M_{\mathrm{do}(X_T = \xi_T)}$ and $\tilde{M}_{\mathrm{do}(X_T = \xi_T)}$ are interventionally equivalent w.r.t. $O \cup T$, that is $\mathcal{X}_O = \tilde{\mathcal{X}}_O$, $\mathcal{X}_{J \cap \tilde{J}} = \tilde{\mathcal{X}}_{J \cap \tilde{J}}$, and for all $\tilde{T} \subseteq O \cup T$:

$$P_{M_{\mathrm{do}(X_T = \xi_T)}}(X_{O \cup T} \mid \mathrm{do}(X_{J \setminus T}, X_{\tilde{T}})) = P_{\tilde{M}_{\mathrm{do}(X_T = \xi_T)}}(X_{O \cup T} \mid \mathrm{do}(X_{\tilde{J} \setminus T}, X_{\tilde{T}})).$$

Note that $\mathrm{do}(X_T = \xi_T)$ can be obtained by $\mathrm{do}(T)$ followed by evaluating in $X_T = \xi_T$:

$$P_{M_{\mathrm{do}(X_T = \xi_T)}}(X_{O \cup T} \mid \mathrm{do}(X_{J \setminus T \cup \tilde{T}})) = P_M(X_{O \cup T} \mid \mathrm{do}(X_{J \setminus T \cup \tilde{T}}), \mathrm{do}(X_T = \xi_T))$$

$$P_{\tilde{M}_{\mathrm{do}(X_T = \xi_T)}}(X_{O \cup T} \mid \mathrm{do}(X_{\tilde{J} \setminus T \cup \tilde{T}})) = P_{\tilde{M}}(X_{O \cup T} \mid \mathrm{do}(X_{\tilde{J} \setminus T \cup \tilde{T}}), \mathrm{do}(X_T = \xi_T))$$

So we want that

$$P_M(X_O \mid \mathrm{do}(X_{J \cup T \cup \tilde{T}})) = P_{\tilde{M}}(X_O \mid \mathrm{do}(X_{\tilde{J} \cup T \cup \tilde{T}}))$$

which we know holds for all $T \cup \tilde{T} \subseteq O \cup (J \cap \tilde{J})$ by the interventional equivalence of M and $\tilde{M}$ w.r.t. O.

3. Let $T \subseteq O \cup (J \cap \tilde{J})$ and $Q_T \in \mathcal{P}(\mathcal{X}_T)$. We can use again that

$$P_M(X_O \mid \mathrm{do}(X_{J \cup T \cup \tilde{T}})) = P_{\tilde{M}}(X_O \mid \mathrm{do}(X_{\tilde{J} \cup T \cup \tilde{T}}))$$

for all $T \cup \tilde{T} \subseteq O \cup (J \cap \tilde{J})$ by the interventional equivalence of M and $\tilde{M}$ w.r.t. O. Note that $\mathrm{do}(X_T \sim Q_T)$ can be obtained by $\mathrm{do}(T)$ followed by putting distribution Q_T on X_T:

$$P_{M_{\mathrm{do}(X_T \sim Q_T)}}(X_{O \cup T} \mid \mathrm{do}(X_{J \setminus T \cup \tilde{T}})) = P_M(X_{O \cup T} \mid \mathrm{do}(X_{J \setminus T \cup \tilde{T}}), \mathrm{do}(X_T)) \circ Q_T(X_T)$$

$$P_{\tilde{M}_{\mathrm{do}(X_T \sim Q_T)}}(X_{O \cup T} \mid \mathrm{do}(X_{\tilde{J} \setminus T \cup \tilde{T}})) = P_{\tilde{M}}(X_{O \cup T} \mid \mathrm{do}(X_{\tilde{J} \setminus T \cup \tilde{T}}), \mathrm{do}(X_T)) \circ Q_T(X_T)$$

Hence, for all $\tilde{T} \subseteq O \cup T$:

$$P_{M_{\mathrm{do}(X_T \sim Q_T)}}(X_{O \cup T} \mid \mathrm{do}(X_{J \setminus T}), \mathrm{do}(X_{\tilde{T}})) = P_{\tilde{M}_{\mathrm{do}(X_T \sim Q_T)}}(X_{O \cup T} \mid \mathrm{do}(X_{\tilde{J} \setminus T}), \mathrm{do}(X_{\tilde{T}}))$$

which implies that $M_{\mathrm{do}(X_T \sim Q_T)}$ is interventionally equivalent to $\tilde{M}_{\mathrm{do}(X_T \sim Q_T)}$ w.r.t. $O \cup T$.

4. For $T \subseteq (J \cap \tilde{J})$, the operation of adding intervention variables $\mathrm{do}(I_T)$ only adds indices $I_j := j$ for $j \in T \cap J$. Therefore, we can consider $T \subseteq O$ without loss of generality, employing Proposition 6.4.14.

   For $\tilde{T} \subseteq O$, we get:

$$\begin{aligned}
&P_{(M_{\mathrm{do}(I_T)})_{\mathrm{do}(\tilde{T})}}(X_{V \setminus \tilde{T}} \mid \mathrm{do}(X_{J \cup \tilde{T}}, X_{I_T} = x_{I_T})) \\
&= P_{\left((M_{\mathrm{do}(\tilde{T} \setminus T)})_{\mathrm{do}(I_T)}\right)_{\mathrm{do}(\tilde{T} \cap T)}}(X_{V \setminus \tilde{T}} \mid \mathrm{do}(X_{J \cup \tilde{T}}, X_{I_T} = x_{I_T})) \\
&= P_{\left((M_{\mathrm{do}(\tilde{T} \setminus T)})_{\mathrm{do}(I_{T \setminus \tilde{T}})}\right)_{\mathrm{do}(\tilde{T} \cap T)}}(X_{V \setminus \tilde{T}} \mid \mathrm{do}(X_{J \cup \tilde{T}}, X_{I_{T \setminus \tilde{T}}} = x_{I_{T \setminus \tilde{T}}})) \\
&= P_{(M_{\mathrm{do}(\tilde{T})})_{\mathrm{do}(I_{T \setminus \tilde{T}})}}(X_{V \setminus \tilde{T}} \mid \mathrm{do}(X_{J \cup \tilde{T}}, X_{I_{T \setminus \tilde{T}}} = x_{I_{T \setminus \tilde{T}}})) \\
&= P_M(X_{V \setminus \tilde{T}} \mid \mathrm{do}(X_{J \cup \tilde{T}}, X_{T_2 \setminus \tilde{T}} = x_{I_{T_2 \setminus \tilde{T}}}))
\end{aligned}$$

   Similar reasoning for $\tilde{M}$ gives:

$$\begin{aligned}
&P_{(\tilde{M}_{\mathrm{do}(I_T)})_{\mathrm{do}(\tilde{T})}}(X_{\tilde{V} \setminus \tilde{T}} \mid \mathrm{do}(X_{\tilde{J} \cup \tilde{T}}, X_{I_T} = x_{I_T})) \\
&= P_{\tilde{M}}(X_{\tilde{V} \setminus \tilde{T}} \mid \mathrm{do}(X_{\tilde{J} \cup \tilde{T}}, X_{T_2 \setminus \tilde{T}} = x_{I_{T_2 \setminus \tilde{T}}}))
\end{aligned}$$

   Since M and $\tilde{M}$ are interventionally equivalent w.r.t. O,

$$\begin{aligned}
&P_M(X_{O \setminus \tilde{T}} \mid \mathrm{do}(X_{J \cup \tilde{T}}, X_{T_2 \setminus \tilde{T}} = x_{I_{T_2 \setminus \tilde{T}}})) \\
&= P_{\tilde{M}}(X_{O \setminus \tilde{T}} \mid \mathrm{do}(X_{\tilde{J} \cup \tilde{T}}, X_{T_2 \setminus \tilde{T}} = x_{I_{T_2 \setminus \tilde{T}}}))
\end{aligned}$$

   Combining this gives:

$$P_{(M_{\mathrm{do}(I_T)})_{\mathrm{do}(\tilde{T})}}(X_{O \setminus \tilde{T}} \mid \mathrm{do}(X_{J \cup \tilde{T}}, X_{I_T})) = P_{(\tilde{M}_{\mathrm{do}(I_T)})_{\mathrm{do}(\tilde{T})}}(X_{O \setminus \tilde{T}} \mid \mathrm{do}(X_{\tilde{J} \cup \tilde{T}}, X_{I_T}))$$

   Hence, $M_{\mathrm{do}(I_T)}$ and $\tilde{M}_{\mathrm{do}(I_T)}$ are interventionally equivalent w.r.t. O.

□

6.8. Marginalizations

When modeling a system, we sometimes want to "hide" details of a subsystem. The following operation on iSCMs that we call "marginalization" is a causal analogue of the marginalization of probability distributions. The computer program analogy of the marginalization operation is to hide details within a subroutine. Intuitively, a marginalization of an iSCM over a subset of endogenous variables L is obtained by first solving a subsystem (the structural equations corresponding to the endogenous variables in L) followed by substituting the solution function of the subsystem into the remaining structural equations (corresponding to the endogenous variables in $V \setminus L$).

Definition 6.8.1. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM and $L \subseteq V$. Suppose that M is essentially uniquely solvable w.r.t. L and let $g^{[L]} : \mathcal{X}_J \times \mathcal{X}_{(V\setminus L)} \times \mathcal{X}_W \to \mathcal{X}_L$ be a corresponding partial solution function of M w.r.t. L. Then we call $M_{\setminus L} = (J, V \setminus L, W, \mathcal{X}_J \times \mathcal{X}_{V\setminus L} \times \mathcal{X}_W, P, f^{\setminus L})$ with*

$$f^{\setminus L}(x_J, x_{V\setminus L}, x_W) := f_{V\setminus L}(x_J, x_{V\setminus L}, g^{[L]}(x_J, x_{V\setminus L}, x_W), x_W)$$

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

If $g^{[L]}$ and $\tilde{g}^{[L]}$ are two partial solution functions of M w.r.t. L, then their corresponding marginalizations of M may not be identical, but they are equivalent. Even more:

Proposition 6.8.2. *If $M \equiv \tilde{M}$ are equivalent iSCMs, $g^{[L]}$ is a partial solution function of M w.r.t. L, and $\tilde{g}^{[L]}$ is a partial solution function of $\tilde{M}$ w.r.t. L, then the marginalizations $M_{\setminus L}$ (corresponding to $g^{[L]}$) and $\tilde{M}_{\setminus L}$ (corresponding to $\tilde{g}^{[L]}$) are equivalent.*

Proof. Since $\tilde{g}^{[L]}$ is a partial solution function of $\tilde{M}$ w.r.t. L, and $M \equiv \tilde{M}$, $\tilde{g}^{[L]}$ is also a partial solution function of M w.r.t. L. Since M is essentially uniquely solvable w.r.t. L, $g^{[L]} = \tilde{g}^{[L]}$ M-a.s.. Therefore, for $v \in V \setminus L$:

$$[x_v = f_v(x_J, x_{V\setminus L}, g^{[L]}(x), x_W) \iff x_v = \tilde{f}_v(x_J, x_{V\setminus L}, \tilde{g}^{[L]}(x), x_W)] \qquad M\text{-a.s.}$$

This implies, for $v \in V \setminus L$:

$$[x_v = f_v^{\setminus L}(x) \iff x_v = \tilde{f}_v^{\setminus L}(x)] \qquad M_{\setminus L}\text{-a.s..}$$

Indeed, if a property $\pi(x_J, x_W, x_{V\setminus L})$ holds M-a.s., then it holds $M_{\setminus L}$-a.s.. □

This shows that even though marginalizations are not unique at the iSCM level, they are unique at the level of equivalence classes of iSCMs.

For simple iSCMs, marginalizations are obviously defined over all subsets $L \subseteq V$.

Notation 6.8.3. *We will also use the notation $M_O := M_{\setminus(V\setminus O)}$ for a subset $O \subseteq V$ of endogenous variables when we want to emphasize which endogenous variables remain after marginalization.*

Marginalization preserves (partial) essentially unique solvability.

Lemma 6.8.4. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM and $L \subseteq V$ such that M is essentially uniquely solvable w.r.t. L. Suppose M is also essentially uniquely solvable w.r.t. $O \subseteq V$ with $L \subseteq O$, and partial solution function $g^{[O]}$. Then its marginalization $M_{\setminus L}$ is essentially uniquely solvable w.r.t. $O \setminus L$, with partial solution function $g^{[O]}_{O\setminus L} = \mathrm{pr}_{O\setminus L} \circ g^{[O]}$ where $\mathrm{pr}_{O\setminus L} : \mathcal{X}_V \to \mathcal{X}_{O\setminus L} : x \mapsto x_{O\setminus L}$ is the canonical projection on $O \setminus L$.*

Proof. Denote by $f^{\setminus L}$ the causal mechanism of the marginalization $M_{\setminus L}$ constructed from the partial solution function $g^{[L]} : \mathcal{X}_J \times \mathcal{X}_{(V \setminus L)} \times \mathcal{X}_W \to \mathcal{X}_L$. From the essentially unique solvability of M w.r.t. O, and the essentially unique solvability of M w.r.t. L, we derive that M-a.s.:

$$\begin{aligned}
&\begin{cases} x_L = g^{[O]}_L(x_J, x_{V\setminus O}, x_W) \\ x_{O\setminus L} = g^{[O]}_{O\setminus L}(x_J, x_{V\setminus O}, x_W) \end{cases} \iff x_O = g^{[O]}(x_J, x_{V\setminus O}, x_W) \iff x_O = f_O(x) \\
\iff &\begin{cases} x_L = f_L(x) \\ x_{O\setminus L} = f_{O\setminus L}(x) \end{cases} \iff \begin{cases} x_L = g^{[L]}(x_J, x_{V\setminus L}, x_W) \\ x_{O\setminus L} = f_{O\setminus L}(x_J, x_{V\setminus L}, x_L, x_W) \end{cases} \\
\iff &\begin{cases} x_L = g^{[L]}(x_J, x_{V\setminus L}, x_W) \\ x_{O\setminus L} = f_{O\setminus L}(x_J, x_{V\setminus L}, g^{[L]}(x_J, x_{V\setminus L}, x_W), x_W) \end{cases} \\
\iff &\begin{cases} x_L = g^{[L]}(x_J, x_{V\setminus L}, x_W) \\ x_{O\setminus L} = f^{\setminus L}_{O\setminus L}(x_J, x_{V\setminus L}, x_W). \end{cases}
\end{aligned}$$

Hence:

$$x_{O\setminus L} = g^{[O]}_{O\setminus L}(x_J, x_{V\setminus O}, x_W) \iff x_{O\setminus L} = f^{\setminus L}_{O\setminus L}(x_J, x_{V\setminus L}, x_W) \quad M\text{-a.s..}$$

A property $\pi(x_J, x_W, x_{V\setminus L})$ that holds M-a.s., also holds $M_{\setminus L}$-a.s., hence:

$$x_{O\setminus L} = g^{[O]}_{O\setminus L}(x_J, x_{V\setminus O}, x_W) \iff x_{O\setminus L} = f^{\setminus L}_{O\setminus L}(x_J, x_{V\setminus L}, x_W) \quad M^{\setminus L}\text{-a.s..}$$

Therefore, the marginalized iSCM $M_{\setminus L}$ is essentially uniquely solvable w.r.t. $O \setminus L$, with essentially unique partial solution function $g^{[O]}_{O\setminus L} = \mathrm{pr}_{O\setminus L} \circ g^{[O]}$.[41] □

Remark 6.8.5. *This also directly implies that if M is essentially uniquely solvable and its marginalization $M_{\setminus L}$ over $L \subseteq V$ is defined, the Markov kernel of the marginalization is obtained by marginalizing the original Markov kernel:*

$$P_{M_{\setminus L}}(X_{V\setminus L} \mid \mathrm{do}(X_J)) = P_M(X_{V\setminus L} \mid \mathrm{do}(X_J)).$$

This explains the name 'marginalization'.

It is now easy to see that simplicity is preserved under marginalization.

Proposition 6.8.6. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. For any $L \subseteq V$, its marginalization $M_{\setminus L}$ is also simple.*

Proof. Let $\tilde{T} \subseteq V \setminus L$. From Lemma 6.8.4 it follows that since M is essentially uniquely solvable w.r.t. $\tilde{T} \cup L$, $M_{\setminus L}$ is essentially uniquely solvable w.r.t. $\tilde{T}$. Hence $M_{\setminus L}$ is simple. □

[41] The converse does not necessarily hold: even if the marginalized iSCM $M_{\setminus L}$ is essentially uniquely solvable w.r.t. $O \setminus L$, the iSCM M may not be essentially uniquely solvable w.r.t. O: $M^{\setminus L}$-a.s. does not imply M-a.s..

Under certain conditions, interventions and marginalization commute (i.e., it does not matter in which order we apply them).

Proposition 6.8.7. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. For $L \subseteq V$ such that M is essentially uniquely solvable w.r.t. L, and an intervention target $T \subseteq V \cup J$ such that $L \cap T = \emptyset$, an intervention targeting T (hard, soft, or adding intervention variables) commutes with marginalizing out L.*

1. *$(M_{\mathrm{do}(T\ldots)})_{\setminus L} \equiv (M_{\setminus L})_{\mathrm{do}(T\ldots)}$ for $T \subseteq (V \setminus L) \cup J$ (hard interventions of all three types);*

2. *$(M_{\mathrm{do}(T\leftarrow\hat{f}_T)})_{\setminus L} \equiv (M_{\setminus L})_{\mathrm{do}(T\leftarrow\hat{f}_T)}$ for $T \subseteq V \setminus L$ and $\hat{f}_T : \mathcal{X}_J \times \mathcal{X}_W \times \mathcal{X}_V \to \mathcal{X}_T$ a measurable function (soft interventions);*

3. *$(M_{\mathrm{do}(I_T)})_{\setminus L} \equiv (M_{\setminus L})_{\mathrm{do}(I_T)}$ for $T \subseteq (V \setminus L) \cup J$ (adding intervention variables);*

Proof. Note that each intervention preserves the causal mechanism components $f_{V\setminus L}$ of the variables not targeted by the intervention. Hence, we may use the same partial solution function w.r.t. L for both M and $M_{\mathrm{do}(\ldots)}$ to construct their respective marginalizations. This even gives identities. If different partial solution functions are used in the construction of the marginalizations, then we still obtain equivalence. □

We will show that for simple iSCMs, the marginalization operation preserves the causal semantics on the remaining variables. A key step is the following proposition, which gives conditions under which it does not matter whether we marginalize at once or in steps.

Proposition 6.8.8. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM and $L_1, L_2 \subseteq V$ such that $L_1 \cap L_2 = \emptyset$. If M is essentially uniquely solvable w.r.t. L_1, and $(M_{\setminus L_1})$ is essentially uniquely solvable w.r.t. L_2, then M is essentially uniquely solvable w.r.t. $L_1 \cup L_2$, and in that case it does not matter if we first marginalize over L_1 and then L_2, or both at once, i.e:*

$$(M_{\setminus L_1})_{\setminus L_2} \equiv M_{\setminus (L_1 \cup L_2)}.$$

Proof. Write $K_1 = V \setminus L_1$. Let $g^{[L_1]} : \mathcal{X}_{J \cup K_1} \times \mathcal{X}_W \to \mathcal{X}_{L_1}$ be the essentially unique solution function of M w.r.t. L_1:

$$x_{L_1} = g^{[L_1]}(x_J, x_{K_1}, x_W) \iff x_{L_1} = f_{L_1}(x) \quad M\text{-a.s.}$$

Let $\tilde{f} : \mathcal{X}_J \times \mathcal{X}_{K_1} \times \mathcal{X}_W \to \mathcal{X}_{K_1}$ with

$$\tilde{f}(x_J, x_{K_1}, x_W) = f_{K_1}(x_J, x_{K_1}, g^{[L_1]}(x_J, x_{K_1}, x_W), x_W)$$

be the causal mechanism of the marginal iSCM $M_{\setminus L_1}$.

If $M_{\setminus L_1}$ is uniquely solvable w.r.t. L_2, it has an essentially unique solution function $\tilde{g}^{[L_2]} : \mathcal{X}_J \times \mathcal{X}_{V \setminus (L_1 \cup L_2)} \times \mathcal{X}_W \to \mathcal{X}_{L_2}$:

$$x_{L_2} = \tilde{g}^{[L_2]}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) \iff x_{L_2} = \tilde{f}_{L_2}(x_J, x_{L_2}, x_{K_1 \setminus L_2}, x_W) \quad M_{\setminus L_1}\text{-a.s.}$$

For a property $\pi(x_J, x_{V\setminus L_1}, x_W)$: π holds M-a.s. if and only if π holds $M_{\setminus L_1}$-a.s., hence

$$x_{L_2} = \tilde{g}^{[L_2]}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) \iff x_{L_2} = f_{L_2}(x_J, x_{K_1}, g^{[L_1]}(x_J, x_{K_1}, x_W), x_W) \quad M\text{-a.s.}$$

Define the function $h : \mathcal{X}_J \times \mathcal{X}_{V\setminus(L_1\cup L_2)} \times \mathcal{X}_W \to \mathcal{X}_{L_1\cup L_2}$ by

$$\begin{aligned} h_{L_1}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) &= g^{[L_1]}(x_J, x_{K_1\setminus L_2}, \tilde{g}^{[L_2]}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W), x_W) \\ h_{L_2}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) &= \tilde{g}^{[L_2]}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) \end{aligned}$$

Then, M-a.s.:

$$\begin{aligned} &\begin{cases} x_{L_1} &= f_{L_1}(x) \\ x_{L_2} &= f_{L_2}(x) \end{cases} \\ \iff &\begin{cases} x_{L_1} &= g^{[L_1]}(x_J, x_{K_1}, x_W) \\ x_{L_2} &= f_{L_2}(x_J, x_{K_1}, x_{L_1}, x_W) \end{cases} \\ \iff &\begin{cases} x_{L_1} &= g^{[L_1]}(x_J, x_{K_1}, x_W) \\ x_{L_2} &= f_{L_2}(x_J, x_{K_1}, g^{[L_1]}(x_J, x_{K_1}, x_W), x_W) \end{cases} \\ \iff &\begin{cases} x_{L_1} &= g^{[L_1]}(x_J, x_{K_1\setminus L_2}, x_{L_2}, x_W) \\ x_{L_2} &= \tilde{g}^{[L_2]}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) \end{cases} \\ \iff &\begin{cases} x_{L_1} &= g^{[L_1]}(x_J, x_{K_1\setminus L_2}, \tilde{g}^{[L_2]}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W), x_W) \\ x_{L_2} &= \tilde{g}^{[L_2]}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) \end{cases} \\ \iff &x_{L_1\cup L_2} = h(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) \end{aligned}$$

where in the first equivalence we used the essentially unique solvability of M w.r.t. L_1, in the second equivalence we used substitution, in the third equivalence we used the essentially unique solvability of $M_{\setminus L_1}$ w.r.t. $[L_2]$, in the fourth equivalence we used substitution again, and in the fifth equivalence we used the definition of h. Hence,

$$x_{L_1\cup L_2} = f_{L_1\cup L_2}(x) \iff x_{L_1\cup L_2} = h(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) \qquad M\text{-a.s.}$$

Therefore, h is the essentially unique solution function for M w.r.t. $L_1 \cup L_2$, which must therefore be essentially uniquely solvable w.r.t. $L_1 \cup L_2$. By checking the definition, one sees that $(M_{\setminus L_1})_{\setminus L_2} = M_{\setminus(L_1\cup L_2)}$. Because of the ambiguity in the choice of the partial solution functions, we conclude $(M_{\setminus L_1})_{\setminus L_2} \equiv M_{\setminus(L_1\cup L_2)}$. □

Corollary 6.8.9. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. For $L_1, L_2 \subseteq V$ with $L_1 \cap L_2 = \emptyset$,*

$$(M_{\setminus L_1})_{\setminus L_2} \equiv (M_{\setminus L_2})_{\setminus L_1} \equiv M_{\setminus(L_1\cup L_2)}.$$

These commutation relations and compatibilities now allow us to give a straightforward proof that the causal semantics are preserved under marginalization. While this holds generally for SCMs [BFPM21], we will here prove this for simple iSCMs.

Theorem 6.8.10. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM, $L \subseteq V$, and $M_{\setminus L}$ its marginalization over L. Then M and $M_{\setminus L}$ are observably and interventionally equivalent w.r.t. $V \setminus L$.*

Proof. Write $O = V \setminus L$. We first show that the marginal Markov kernels $P_M(X_O \mid \mathrm{do}(X_J))$ and $P_{M_{\setminus L}}(X_O \mid \mathrm{do}(X_J))$ are the same. The former is obtained as:

$$P_M(X_O \mid \mathrm{do}(X_J)) = (\mathrm{pr}_O \circ g)_*(P) = (g_O)_*(P),$$

where $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ is the essentially unique solution function of M and $\mathrm{pr}_O : \mathcal{X}_{V \cup W} \to \mathcal{X}_O$ is the canonical projection on the O components. The latter is obtained as:

$$P_{M_{\setminus L}}(X_O \mid \mathrm{do}(X_J)) = (g_O)_*(P)$$

since by Lemma 6.8.4, g_O is the (essentially unique) solution function of $M_{\setminus L}$. This means that both push-forwards are identical.

Let $T \subseteq O$. Then $(M_{\setminus L})_{\mathrm{do}(T)} \equiv (M_{\mathrm{do}(T)})_{\setminus L}$ by Proposition 6.8.7. The observable equivalence of $M_{\mathrm{do}(T)}$ and $(M_{\mathrm{do}(T)})_{\setminus L}$ w.r.t. $O \setminus T$ hence implies the observable equivalence of $M_{\mathrm{do}(T)}$ and $(M_{\setminus L})_{\mathrm{do}(T)}$ w.r.t. $O \setminus T$ (using Proposition 6.7.6). Since this holds for all $T \subseteq O$, M and $M_{\setminus L}$ are interventionally equivalent w.r.t. O. □

So the marginalization operation indeed effectively hides the details of a subsystem, while preserving the causal semantics on the remaining part.

6.9. Graphical Representations

We can represent the qualitative structure of an iSCM by means of graphs. The directed edges in such a graph will express the following "parent"-relation that captures *functional dependencies* in the structural equations / causal mechanisms of the iSCM.

Definition 6.9.1. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. For $i \in J \cup V \cup W$ and $j \in V$, we say that i* ***is a parent of*** *j* ***according to*** *M if there does not exist a measurable function $\tilde{f}_j : \mathcal{X}_{(J \cup V \cup W) \setminus \{i\}} \to \mathcal{X}_j$ such that*

$$x_j = f_j(x) \iff x_j = \tilde{f}_j(x_{\setminus i}) \qquad M\text{-a.s.},$$

where $x_{\setminus i}$ is shorthand for $x_{(J \cup V \cup W) \setminus \{i\}}$.

In words, i is a parent of j if the solutions of the structural equation for j essentially depend on the value of variabe i. By definition, exogenous (input and random) variables have no parents.

Using directed edges to encode the parent-relationship, we define two graphical representations. The more fine-grained representation represents all the variables as nodes. We will also frequently use a coarser representation that represents only the endogenous variables and the exogenous input variables as nodes (but not the exogenous random variables).

Definition 6.9.2. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. The CDG $(J, V \cup W, E)$ with input nodes J, output nodes $V \cup W$, and directed edges*

$$E = \{i \longrightarrow j : i \in J \cup W \cup V, j \in V : i \text{ is parent of } j \text{ according to } M\}$$

is called the ***graph of the iSCM*** *and will be denoted as $G^+(M)$.*

The marginalized graph $(G^+(M))^{\setminus W}$ in which all exogenous random nodes have been marginalized out (via Definition 3.2.18), will be called the ***causal graph of the iSCM*** *and denoted as $G(M)$.*[42]

Remark 6.9.3. *The causal graph $G(M) = (J, V, E, L)$ has input nodes J, output nodes V, directed edges*

$$E = \{i \longrightarrow j : i \in J \cup V, j \in V : i \text{ is parent of } j \text{ according to } M\}$$

and bidirected edges

$$L = \{j \longleftrightarrow k : j \in V, k \in V, j \neq k : j \text{ and } k \text{ have a common parent in according to } M\}.$$

In other words, $G(M)$ is obtained from $G^+(M)$ by replacing the nodes representing the exogenous random variables and their outgoing directed edges with bidirected edges, i.e., any pattern $\longleftarrow w \longrightarrow$ with $w \in W$ is replaced by $\longleftrightarrow$.

Note that observably equivalent iSCMs may have different graphs, and even interventionally equivalent iSCMs may have different graphs.

Example 6.9.4. *Consider the SCM M with endogenous variables X_1, X_2, X_3 with co-domains $\{-1, 1\}$, $\{-1, 1\}$, $\{-2, 0, 2\}$, respectively, and structural equations*

$$\begin{aligned} X_1 &= X_A \\ X_2 &= X_1 X_B \\ X_3 &= X_2 + X_B, \end{aligned}$$

with independent exogenous random variables $X_A, X_B \sim \mathrm{Uni}(\{-1, 1\})$. Its graph $G^+(M)$ and its causal graph $G(M)$ are depicted in Figure 17 (left top and bottom, respectively).

Consider also the SCM $\tilde{M}$ with endogenous variables X_1, X_2, X_3 with co-domains $\{-1, 1\}$, $\{-1, 1\}$, $\{-2, 0, 2\}$, respectively, and structural equations

$$\begin{aligned} X_1 &= X_A \\ X_2 &= X_C \\ X_3 &= X_2 + X_1 X_C, \end{aligned}$$

with independent exogenous random variables $X_A, X_C \sim \mathrm{Uni}(\{-1, 1\})$. Its graph $G^+(\tilde{M})$ and causal graph $G(\tilde{M})$ are depicted in Figure 17 (right top and bottom, respectively).

[42] The reason we use the adjective "causal" is to emphasize that every node in the causal graph has a causal interpretation, as the corresponding variable can be targeted by a hard intervention.

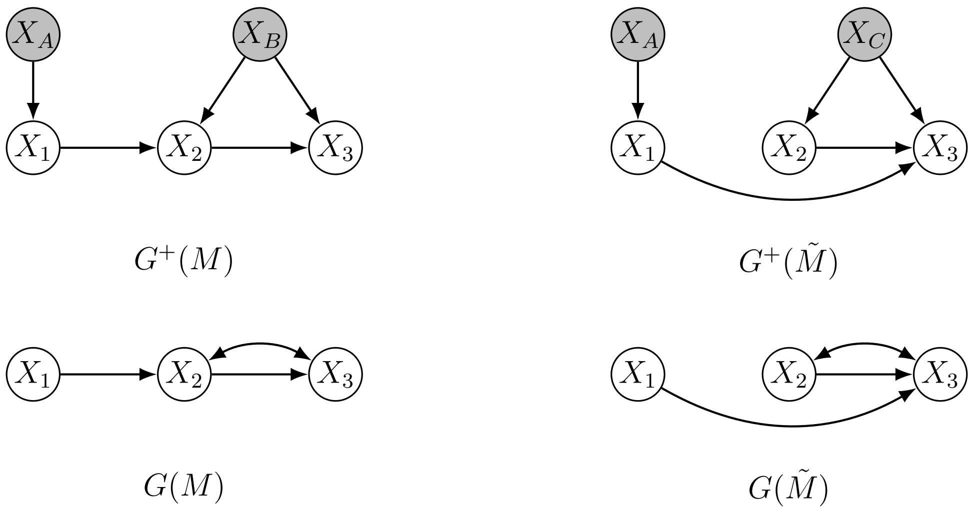


Figure 17: The graphs of the interventionally equivalent SCMs M (left) and $\tilde{M}$ (right) corresponding to Example 6.9.4. The corresponding causal graphs (that do not represent exogenous random variables as nodes) are shown in the bottom row.

$\tilde{M}$ was obtained from M by making a change of variables $x_C = x_1 x_B$. One can check that $\tilde{M}$ is interventionally equivalent to M (with respect to $\{1, 2, 3\}$), even though its the graph $G^+(M)$ differs from the graph $G^+(\tilde{M})$, and its causal graph $G(M)$ differs from the causal graph $G(\tilde{M})$.[43]

Remark 6.9.5. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM and let $G^+ := G^+(M)$ be its graph. Then there exists an equivalent iSCM $\tilde{M} \equiv M$ with causal mechanism $\tilde{f} : \mathcal{X} \to \mathcal{X}_V$ such that for each $v \in V$: $\tilde{f}_v$ is constant in $x_{(J \cup W \cup V) \setminus \mathrm{Pa}^{G^+}(v)}$. Hence, the causal mechanisms of $\tilde{M}$ can be seen as a tuple of functions $(\tilde{f}_v : \mathcal{X}_{\mathrm{Pa}^{G^+}(v)} \to \mathcal{X}_v)_{v \in V}$.*

In the literature, things are often done in the opposite way: given a graph, one considers an SCM such that its causal mechanisms respect the graph structure.

We already provided definitions for hard interventions on graphs (Definition 3.2.1) and for marginalizations (latent projections) of graphs (Definition 3.2.18). The mapping that maps an iSCM to its graph is compatible with the elementary operations on iSCMs and on graphs.

Proposition 6.9.6. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. Then*

- *Hard interventions: for $T \subseteq J \cup V$,*

$$G^+(M_{\mathrm{do}(T)}) = (G^+(M))_{\mathrm{do}(T)} \qquad \text{and} \qquad G(M_{\mathrm{do}(T)}) = (G(M))_{\mathrm{do}(T)}.$$

[43]Even node-splitting interventions would not allow to distinguish the two SCMs.

Figure 18: Graphs without and with a self-cycle in Example 6.10.1.

- *Marginalizations: If M is simple, then for $L \subseteq V$,*

$$G^+(M_{\setminus L}) \subseteq (G^+(M))^{\setminus L}.$$

Proof. The first statement follows by writing out the definitions.

The second statement is somewhat more involved. We will first prove it in case $L = \{\ell\}$ consists of a single node. Let $G^+ := G^+(M)$. By using the definition of the parent relation (repeatedly), we can find a function $\tilde{f}_\ell : \mathcal{X}_{\mathrm{Pa}^{G^+}(\ell)} \to \mathcal{X}_\ell$ such that:

$$x_\ell = f_\ell(x) \iff x_\ell = \tilde{f}_\ell(x_{\mathrm{Pa}^{G^+}(\ell)}) \quad M\text{-a.s.}.$$

Since $\ell \notin \mathrm{Pa}^{G^+}(\ell)$ because M is simple (see Proposition 6.10.2), the essentially unique solution function $\tilde{g}_\ell : \mathcal{X}_{J \cup V \setminus \{\ell\} \cup W} \to \mathcal{X}_\ell$ of $M_{\mathrm{do}(V \setminus \{\ell\})}$ satisfies $\tilde{g}_\ell(x_{\setminus \ell}) = \tilde{f}_\ell(x_{\mathrm{Pa}^{G^+}(\ell)})$ M-a.s., i.e., it only depends on the parents of ℓ. When constructing the marginalized causal mechanism for $M_{\setminus \{\ell\}}$, we substitute $\tilde{g}_\ell(x_{\setminus \ell})$ into the ℓ'th input of the causal mechanism f_j of M, for $j \in V \setminus \{\ell\}$. Since $\tilde{g}_\ell$ only depends on $\mathrm{Pa}^{G^+}(\ell)$, we get that $\mathrm{Pa}^{\tilde{G}^+}(j) \subseteq \mathrm{Pa}^{G^+}(j) \setminus \{\ell\} \cup \mathrm{Pa}^{G^+}(\ell)$, where $\tilde{G}^+ = G^+(M_{\setminus \ell})$. But we also have $\mathrm{Pa}^{(G^+)^{\setminus \{\ell\}}}(j) = \mathrm{Pa}^{G^+}(j) \setminus \{\ell\} \cup \mathrm{Pa}^{G^+}(\ell)$ by definition of the graphical marginalization. Hence $\mathrm{Pa}^{\tilde{G}^+}(j) \subseteq \mathrm{Pa}^{(G^+)^{\setminus \{\ell\}}}$ for all $j \in V \setminus \{\ell\}$, and we have shown that $G^+(M_{\setminus \{\ell\}}) \subseteq (G^+(M))^{\setminus \{\ell\}}$. For the general case, we can make use of induction and the facts that both for graphs and simple iSCMs, we can obtain a marginalization over a subset by repeatedly marginalizing out a single remaining node in the subset, in arbitrary order. Finally, note that $G(M_{\setminus L}) = (G^+(M_{\setminus L}))^{\setminus W} \subseteq ((G^+(M))^{\setminus L})^{\setminus W} = G(M)^{\setminus L}$. □

6.10. Self-Cycles

Our goal was to allow for possible cycles in an iSCM. This means that we may also encounter *self-cycles.*

Example 6.10.1. *Consider an SCM with an endogenous variable $X \in \mathbb{R}$ and an exogenous random variable $W \in \mathbb{R}$.*

If it has structural equation

$$X = X - X^3 + W,$$

then its graph is the one in Figure 18(a). In particular, it does not have a self-cycle at X, since X is not a parent of itself. Indeed, the structural equation is equivalent to

$$X = \sqrt[3]{W},$$

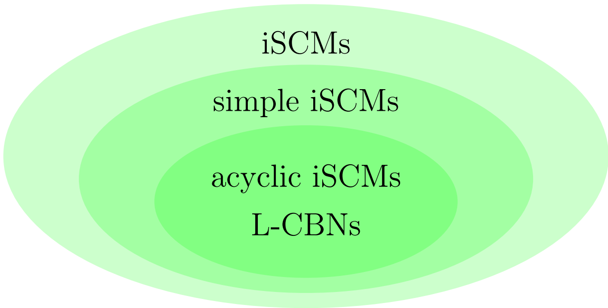


Figure 19: Venn diagram for different causal modeling classes.

where X does not appear on the r.h.s..

On the other hand, if it has structural equation

$$X = X - X^2 + W^2,$$

then its graph is the one in Figure 18(b), which in particular contains a self-cycle at X.

Self-cycles complicate matters because they indicate solvability issues.

To understand why self-cycles are in some sense inevitable, consider an SCM with structural equations

$$\begin{aligned} X_1 &= X_2 \\ X_2 &= X_3 \\ X_3 &= X_1 \end{aligned}$$

Marginalizing out X_2 and X_3 gives an SCM with structural equation

$$X_1 = X_1,$$

which turns the cycle $X_1 \longrightarrow X_3 \longrightarrow X_2 \longrightarrow X_1$ into a self-cycle $X_1 \longrightarrow X_1$. Self-cycles are of no concern, though, when restricting to the class of simple iSCMs.

Proposition 6.10.2. *Let M be an iSCM. For $j \in V$, there is a self-cycle $j \longrightarrow j$ in $G^+(M)$ if and only if M is **not** essentially uniquely solvable w.r.t. j. In particular, graphs of simple iSCMs have no self-cycles.*

Proof. j is not a parent of itself according to M iff there is a function $\tilde{f}_j : \mathcal{X}_{(J \cup V \cup W) \setminus \{j\}} \to \mathcal{X}_j$ such that $x_j = f_j(x) \iff x_j = \tilde{f}_j(x_{\setminus j})$ M-a.s., which holds iff M is essentially uniquely solvable w.r.t. j. $\square$

6.11. Acyclic iSCMs

A subclass of iSCMs that is often considered are acyclic iSCMs.

Definition 6.11.1. *An iSCM M is called **acyclic** if its graph $G^+(M)$ is acyclic.*

Note that this holds if and only if its causal graph $G(M)$ is acyclic. If one models static systems, then using acyclic iSCMs rules out the presence of causal cycles (e.g., feedback loops) in the system. Acyclic iSCMs are a subclass of the more general class of simple iSCMs.

Proposition 6.11.2. *Acyclic iSCMs are simple.*

Proof. We first show that acyclic iSCMs are essentially uniquely solvable. Let M be an acyclic iSCM. Its graph $G^+ := G^+(M)$ is acyclic, and hence has a topological order $<$. Consider f_v, the causal mechanism for $v \in V$. The parents $\mathrm{Pa}^{G^+}(v)$ precede v in the topological order. Without loss of generality, we may assume that each mechanism f_v only depends on its parents (see Remark 6.9.5), and thus consider $f_v : \mathcal{X} \to \mathcal{X}_v$ as a function $f_v : \mathcal{X}_{\mathrm{Pred}_<^{G^+}(v)} \to \mathcal{X}_v$ instead. We can then inductively define the components

$$g_v : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_v : (x_J, x_W) \mapsto f_v(g_{\mathrm{Pred}_<^{G^+}(v)}(x_J, x_W))$$

that together form a solution function $g : \mathcal{X}_{J\cup W} \to X_V$. This construction also exhibits the essential uniqueness of $g : \mathcal{X}_{J\cup W} \to X_V$.

Next consider $M_{\mathrm{do}(T)}$, the intervened iSCM for a hard intervention on M with target $T \subseteq V$. It has graph $G^+(M_{\mathrm{do}(T)}) = (G^+)_{\mathrm{do}(T)}$, whose edges form a subset of the edges of G^+ (but where some output nodes have become input nodes), and hence is also acyclic. Therefore, also $M_{\mathrm{do}(T)}$ is essentially uniquely solvable. One easily sees that the solution function constructed above is also a partial solution function of M w.r.t. $V \setminus T$, which is essentially unique. Since this holds for all targets $T \subseteq V$, we conclude that M is simple. $\square$

Figure 19 shows a Venn diagram to illustrate the relationships between the different classes of causal models that we introduced. We will show that in a precise sense, acyclic iSCMs and L-CBNs are equally expressive.

Definition 6.11.3. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM and let $\tilde{M}$ be an L-CBN with graph $\tilde{G}^+ = (\tilde{J}, \tilde{V} \dot\cup \tilde{U}, \tilde{E})$, spaces $\tilde{\mathcal{X}}_{\tilde{v}}$ for $\tilde{v} \in \tilde{V} \cup \tilde{U} \cup \tilde{J}$, and Markov kernels*

$$P_{\tilde{v}}\left(X_{\tilde{v}} | X_{\mathrm{Pa}^{\tilde{G}^+}(\tilde{v})}\right).$$

*We say that M **is interventionally equivalent to** $\tilde{M}$ **w.r.t.** $O \subseteq V \cap \tilde{V}$ if $\mathcal{X}_O = \tilde{\mathcal{X}}_O$, $\mathcal{X}_{J\cap\tilde{J}} = \tilde{\mathcal{X}}_{J\cap\tilde{J}}$, and for any hard intervention $\mathrm{do}(T)$ with $T \subseteq O$, the intervened Markov kernel $P_M(X_{O\setminus T} | \mathrm{do}(X_{J\cup T}))$ of M equals the intervened Markov kernel $P(X_{O\setminus T} | \mathrm{do}(X_{\tilde{J}\cup T}))$ of $\tilde{M}$.*

Proposition 6.11.4. *i) Given an acyclic iSCM $M = (J, V, W, \mathcal{X}, P, f)$ and a subset $O \subseteq V$, we can construct an L-CBN $\tilde{M}$ with observed output variables $\tilde{V} = O$, latent output variables $\tilde{U} = (V \cup W) \setminus O$, input variables J, graph $G^+(M)$, and spaces $\mathcal{X}_v$ for $v \in V \cup W \cup J$ that is interventionally equivalent to M w.r.t. O.*

ii) Given an L-CBN $\tilde{M} = \left(G^+ = (J,(O,U),E^+),\ \left(P_v(X_v|X_{\mathrm{Pa}^{G^+}(v)})\right)_{v\in O\cup U}\right)$ *with observed variables* O*, we can construct an acyclic iSCM* M *with input variables* J*, endogenous variables* O *and causal graph* $G(M) \subseteq (G^+)^{\setminus U}$ *that is interventionally equivalent to* $\tilde{M}$ *w.r.t.* O*.*

Proof. i) Let $M = (J, V, W, \mathcal{X}, P, f)$ be an acyclic iSCM with graph $G^+(M) = (J, V \cup W, E)$. Define $\tilde{M}$ as the L-CBN with observed output variables $\tilde{V} = O$, latent output variables $\tilde{U} = (V \cup W) \setminus O$, input variables J, graph $G^+ := G^+(M)$, spaces $\mathcal{X}_v$ for $v \in V \cup W \cup J$, and the following Markov kernels. For $v \in V$, we write its structural equation as:

$$x_v = f_v(x_{\mathrm{Pa}^{G^+}(v)})$$

with $f_v : \mathcal{X}_{\mathrm{Pa}^{G^+}(v)} \to \mathcal{X}_v$, where $v \notin \mathrm{Pa}^{G^+}(v)$ because the graph is acyclic. We then define the corresponding (deterministic) Markov kernel

$$P_v\left(X_v|X_{\mathrm{Pa}^{G^+}(v)}\right) := \delta_{f_v}(X_v|X_{\mathrm{Pa}^{G^+}(v)}),$$

encoding the causal mechanisms of M. The Markov kernels for $w \in W$ are defined as:

$$P_w\left(X_w|X_{\mathrm{Pa}^{G^+}(w)}\right) := P_w(X_w),$$

encoding the exogenous distributions of M, where we note that $\mathrm{Pa}^{G^+}(w) = \emptyset$. One can check that this L-CBN $\tilde{M}$ does the job.

ii) Let

$$\tilde{M} = \left(G^+ = (J,(O,U),E^+),\ \left(P_v(X_v|X_{\mathrm{Pa}^{G^+}(v)})\right)_{v\in O\cup U}\right)$$

be an L-CBN. For every $v \in O \dot{\cup} U$, we can write the Markov kernel P_v as the composition of a deterministic one and a uniform distribution $P_{\bar{v}}(X_{\bar{v}})$ on $\mathcal{X}_{\bar{v}} := [0, 1]$ by Remark 2.7.4:

$$P_v(X_v|X_{\mathrm{Pa}^{G^+}(v)}) = \delta(f_v|X_{\bar{v}}, X_{\mathrm{Pa}^{G^+}(v)}) \circ P_{\bar{v}}(X_{\bar{v}})$$

for some measurable function $f_v : \mathcal{X}_{\bar{v}} \times \mathcal{X}_{\mathrm{Pa}^{G^+}(v)} \to \mathcal{X}_v$. Here, we introduced new variables $\bar{v}$ for each $v \in O \dot{\cup} U$.

Define now the iSCM $\bar{M} = \left(J, O \dot{\cup} U, \bar{O} \dot{\cup} \bar{U}, \mathcal{X}_J \times \mathcal{X}_{O \dot{\cup} U} \times \mathcal{X}_{\bar{O} \dot{\cup} \bar{U}}, P, f\right)$ by taking

$$P(X_{\bar{O} \dot{\cup} \bar{U}}) = \bigotimes_{\bar{v}\in \bar{O} \dot{\cup} \bar{U}} P_{\bar{v}}(X_{\bar{v}})$$

and

$$f = (f_v)_{v\in O \dot{\cup} U}.$$

That is, the uniformly distributed random variables $X_{\bar{v}}$ become the exogenous random variables, all independent and uniformly distributed on $[0, 1]$, and the components of the causal mechanism correspond to the deterministic functions used to represent the Markov kernels. The marginalized iSCM $M := \bar{M}_{\setminus U}$ does the job, as one can check.

$\square$

Simple iSCMs are more expressive than acyclic iSCMs because they can model (sufficiently weak) causal cycles. iSCMs in general are even more expressive because they can also model stronger cycles that not necessarily lead to unique solvability under any hard intervention, but this generality comes with a substantially increased complexity of the theory and interpretability. Simple iSCMs form a "sweet spot" in the sense that they allow cyclic relationships yet their theory is not much more complicated than that of acyclic iSCMs: the main difference consists in replacing d-separation with σ-separation.

Exercise 6.11.5. *Show that all simple iSCMs with two endogenous binary variables X, Y must be acyclic. Give an example of a cyclic simple iSCM with two endogenous variables X, Y that take on values in $\{0, 1, 2\}$.*

Even iSCMs may not be the ultimate way of modeling cyclic causal systems. Indeed, for such systems, it might be that the conceptual notion of interventions *targeting variables* is misguided in general, and perhaps should be replaced by the notion of intervening on *functional constraints* [BvDM21].

6.12. Examples

In many systems occurring in the real world, feedback loops between observed variables are present. Such systems can often be described by a system of (random) differential equations. The equilibrium states of such systems can sometimes be causally modelled by an iSCM [BM18].

For illustration purposes we provide two examples, the first consisting of interacting masses that are attached to springs that can be described at equilibrium with a simple iSCM, the second being the famous price-supply-demand model that has been very popular in econometrics, and which corresponds to a non-simple SCM at equilibrium.

Example 6.12.1 (Damped coupled harmonic oscillator)**.** *Consider a one-dimensional system of d masses $m_i \in \mathbb{R}$ $(i = 1, \ldots, d)$ with positions Q_i. The masses are coupled by springs, with spring constants $k_i > 0$ $(i = 0, \ldots, d)$ and equilibrium lengths $\ell_i > 0$ $(i = 0, \ldots, d-1)$, under influence of friction with friction coefficients $b_i > 0$ $(i = 1, \ldots, d)$. The endpoints are considered fixed at positions $Q_0 < Q_{d+1}$ (see Figure 20 (top)). From elementary physics, we know that the equations of motion of this system are given by the following differential equations*

$$\frac{d^2 Q_i}{dt^2} = \frac{k_i}{m_i}(Q_{i+1} - Q_i - \ell_i) + \frac{k_{i-1}}{m_i}(Q_{i-1} - Q_i + \ell_{i-1}) - \frac{b_i}{m_i}\frac{dQ_i}{dt} \qquad i = 1, \ldots, d.$$

The dynamics of the masses, in terms of the position Q_i, velocity $\frac{dQ_i}{dt}$ and acceleration $\frac{dQ_i^2}{dt^2}$, is described by a single and separate equation of motion for each mass. Under

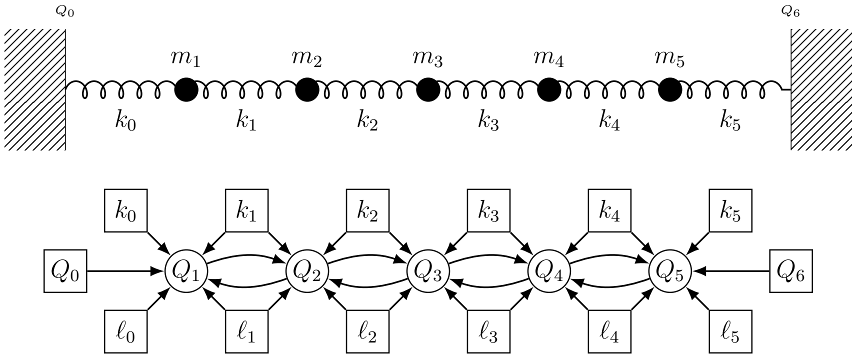


Figure 20: Damped coupled harmonic oscillator (top) and the graph of the iSCM that describes the positions of the masses at equilibrium (bottom) of Example 6.12.1 for $d = 5$, where the spring lengths and constants are considered as exogenous input variables.

friction, i.e., $b_i > 0$ ($i = 1, \dots, d$), there is a unique equilibrium position, where the sum of forces vanishes for each mass. If one moves one or several masses out of their equilibrium positions and releases them, then the masses will start to oscillate, but eventually these oscillations dampen out and the masses converge to their unique equilibrium position. At equilibrium (i.e., for $t \to \infty$) the velocity $\frac{dQ_i}{dt}$ and acceleration $\frac{d^2Q_i}{dt^2}$ of the masses vanish (i.e., $\frac{dQ_i}{dt}, \frac{d^2Q_i}{dt^2} \to 0$), and thus the following equation holds at equilibrium

$$0 = \frac{k_i}{m_i}(Q_{i+1} - Q_i - \ell_i) + \frac{k_{i-1}}{m_i}(Q_{i-1} - Q_i + \ell_{i-1})$$

for each mass ($i = 1, \dots, d$). By solving each of these equations w.r.t. Q_i, we obtain that the equilibrium positions Q_i of the masses are given by

$$Q_i = \frac{k_i(Q_{i+1} - \ell_i) + k_{i-1}(Q_{i-1} + \ell_{i-1})}{k_i + k_{i-1}}.$$

By considering the ℓ_i, k_i and Q_0 and Q_{d+1} as exogenous input variables, and the Q_i ($i = 1, \dots, d$) as endogenous variables, we arrive at an iSCM with causal mechanism

$$f_i(q, \ell, k) = \frac{k_i(q_{i+1} - \ell_i) + k_{i-1}(q_{i-1} + \ell_{i-1})}{k_i + k_{i-1}}.$$

for $i = 1, \dots, d$. Its graph is depicted in Figure 20 (bottom). This iSCM allows us to describe the equilibrium behavior of the system under perfect intervention. For example, when forcing the mass j to a fixed position $Q_j = \xi_j$ with $0 \le \xi_j \le L$, the equilibrium positions of the masses correspond to the solutions of the intervened model $M_{\mathrm{do}(Q_j=\xi_j)}$.

Exercise 6.12.2. *Prove that the iSCM that describes the equilibrium states of a damped coupled harmonic oscillator is simple (see also Proposition 6.6.14). Hint: you can use*

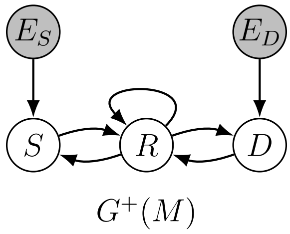


Figure 21: The graph of the SCM M of Example 6.12.3.

that the determinant of a tridiagonal matrix of the following form is given by the expression on the r.h.s.:

$$\det \begin{pmatrix} k_0 + k_1 & -k_1 & & & & \\ -k_1 & k_1 + k_2 & -k_2 & & & \\ & -k_2 & k_2 + k_3 & \ddots & & \\ & & \ddots & \ddots & -k_{d-1} \\ & & & -k_{d-1} & k_{d-1} + k_d \end{pmatrix} = \sum_{i=0}^{d} \prod_{\substack{j=0 \\ j \neq i}}^{d} k_j$$

Next, we show that a well-known equilibrium model from economics can be described by a (non-simple) SCM. This example illustrates how self-cycles enrich the class of SCMs.

Example 6.12.3 (Price, supply and demand)**.** *Let D denote the demand and S the supply of a quantity of a product. The price of the product is denoted by R. The following system of differential equations describes how the demanded and supplied quantities are determined by the price, and how price adjustments occur in the market:*

$$D = \beta_D R + E_D$$
$$S = \beta_S R + E_S$$
$$\frac{dR}{dt} = D - S,$$

where E_D and E_S are exogenous random influences on the demand and supply respectively, $\beta_D < 0$ is the reciprocal of the slope of the demand curve, and $\beta_S > 0$ is the reciprocal of the slope of the supply curve. At equilibrium, $\frac{dR}{dt} = 0$, and hence the price is determined implicitly by the condition that demanded and supplied quantities should be equal. At equilibrium, hence, we obtain an SCM M with causal mechanism defined by:

$$f_D(d, s, r, e_D, e_S) := \beta_D r + e_D$$
$$f_S(d, s, r, e_D, e_S) := \beta_S r + e_S$$
$$f_R(d, s, r, e_D, e_S) := r + (d - s).$$

Note how we use a self-cycle for r in order to implement the equilibrium equation $d = s$ as the causal mechanism for the price r. Its graph is depicted in Figure 21 (left).

Exercise 6.12.4. *Prove that the SCM M that describes the equilibrium states of the price-supply-demand model is essentially uniquely solvable, but not simple. Consider the following interventions:* $\mathrm{do}(D = \delta)$, $\mathrm{do}(S = \sigma)$, $\mathrm{do}(R = \rho)$, *and all possible combinations thereof. Which of (the combinations of) these interventions give an intervened SCM that is still essentially uniquely solvable? Which of these interventions on the SCM correspond with the equilibrium state of a similarly intervened market dynamics model? Summarizing: could this be a realistic causal equilibrium model of an ideal market, or is there something wrong with it (perhaps due to the self-cycle)?*
(Bonus: can you model the equilibrium with an SCM ***without*** *self-cycles?)*

While the price-supply-demand example shows that not all cyclic SCMs that occur "in the wild" are simple, we have chosen to restrict ourselves mostly to simple iSCMs for this lecture. Generalizations of the theory presented here for simple iSCMs to non-simple ones—but without inputs—are provided in [BFPM21].

We will finish this section by showing that iSCMs are simple if the causal mechanisms are sufficiently weak and smooth.

Proposition 6.12.5. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM with real-valued endogenous variables, that is, $\mathcal{X}_v = \mathbb{R}$ for each $v \in V$. If for each subset $U \subseteq V$, and for all values $x_W \in \mathcal{X}_W$, $x_J \in \mathcal{X}_J$, $x_{V \setminus U} \in \mathcal{X}_{V \setminus U}$, the mapping*

$$\mathcal{X}_U \to \mathcal{X}_U : x_U \mapsto f(x_J, x_{V\setminus U}, x_U, x_W)$$

is Lipschitz continuous with Lipschitz constant $L_U(x_J, x_{V\setminus U}, x_W) < 1$ with respect to some norm $||\cdot||$, then M is simple.

Proof. By definition, M is simple if for all $U \subseteq V$, for all $x_W \in \mathcal{X}_W$, for all $x_J \in \mathcal{X}_J$, for all $x_{V\setminus U} \in \mathcal{X}_{V\setminus U}$, the equation

$$x_U = f(x_J, x_{V\setminus U}, x_U, x_W) \tag{42}$$

has a unique solution for $x_U \in \mathcal{X}_U$. This is a fixed point equation for x_U, and hence it has a unique solution by Banach's fixed point theorem if it is a contraction (Lipschitz continuous with Lipschitz constant < 1 with respect to some norm $||\cdot||$). □

Remark 6.12.6. *This also provides us with a method for sampling from a simple iSCM that satisfies the assumption in Proposition 6.12.5. The solution to equation (42) can be obtained by iterating the updates*

$$x_U^{(n+1)} = f(x_J, x_{V\setminus U}, x_U^{(n)}, x_W)$$

until convergence.

To give a more concrete example: a class of iSCMs that satisfies the contractivity condition is given by neural networks with sufficiently weak weights.

Example 6.12.7. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM with real-valued variables, that is, $\mathcal{X}_k = \mathbb{R}$ for each $k \in V \cup W \cup J$. Suppose that the causal mechanism is of the form*

$$f_u = h\left(\sum_{j \in J} A_{uj} x_j + \sum_{w \in W} A_{uw} x_w + \sum_{v \in V} A_{uv} x_v + b_u\right), \qquad u \in V,$$

with weights $A \in \mathbb{R}^{V \times (V \cup W \cup J)}$, biases $b \in \mathbb{R}^V$ and activation function $h : \mathbb{R} \to \mathbb{R}$.

The conditions in Proposition 6.12.5 are satisfied if the following conditions both hold:

1. $\sup_{x \in \mathbb{R}} |h'(x)| \leq C$ *with* $0 < C < \infty$, *and*
2. $||A_{UU}|| < \frac{1}{C}$ *for every subset* $U \subseteq V$ *of cardinality* $\#(U) \geq 2$, *where* $|| \cdot ||$ *can be one of the matrix norms:* $|| \cdot ||_p$, $p \geq 1$, *or* $|| \cdot ||_\infty$.

Proof. By the mean value theorem, it suffices to show that for every subset $U \subseteq V$ of cardinality $\#(U) \geq 2$ and every value $(x_J, x_W, x_{V \setminus U})$ the partial derivative is bounded:

$$\sup_{x_U \in \mathcal{X}_U} \left\| \frac{\partial f_U}{\partial x_U}(x_J, x_{V \setminus U}, x_U, x_W) \right\| \leq L_U(x_J, x_{V \setminus U}, x_W) < 1$$

for $|| \cdot ||$ some matrix norm. In our case we have:

$$\frac{\partial f_U}{\partial x_U}(x_J, x_{V \setminus U}, x_U, x_W) = \operatorname{diag}(\eta)_{UU} A_{UU}.$$

where η is a vector in $\mathbb{R}^V$ with entries

$$\eta_v = h'(A_v(x_J, x_V, x_W)^\top + b_v).$$

If $|h'(x)| \leq C < \infty$ for all $x \in \mathbb{R}$, and $|| \cdot ||$ is either $|| \cdot ||_p$, $p \geq 1$, or $|| \cdot ||_\infty$, then $|| \operatorname{diag}(\eta)_{UU}|| \leq C$. Since $||A_{UU}|| < \frac{1}{C}$ we get

$$|| \operatorname{diag}(\eta)_{UU} A_{UU}|| \leq || \operatorname{diag}(\eta)_{UU}|| \cdot ||A_{UU}|| =: L_U(x_J, x_{V \setminus U}, x_W) < C\frac{1}{C} = 1.$$

□

Remark 6.12.8. *Note that we can put $C = 1$ for popular activation functions $h(x)$ like* $\tanh(x)$, $\mathrm{ReLU}(x) = \max(0, x)$, $\sigma(x) = \frac{1}{1+\exp(-x)}$, *LeakyRelu, and SoftPlus* $(x) = \ln(1 + e^x)$. *Further note that by using one of these activation functions $h(x)$ and $|| \cdot || = || \cdot ||_\infty$ all the conditions are satisfied if we choose the weights $A_{v,k}$ such that for all $v \in V$:*

$$\sum_{k \in V \cup W \cup J} |A_{v,k}| < 1,$$

and $A_{v,v} = 0$. While this is far from necessary, it is easy to check.

7. Markov Property for iSCMs and its consequences

7.1. Acyclifications

By making use of the ‘acyclification’, we can extend the global Markov property for L-CBNs to a global Markov property for simple iSCMs. The difference is that the Markov property for iSCMs is formulated in terms of σ-separation rather than d-separation. With the help of this Markov property, we can derive a theory for simple iSCMs that is very similar to that of L-CBNs, with a do-calculus and adjustment. In Section 3.5, we defined acyclifications of a CDMG. We can also define an operation with the same name on iSCMs.

Definition 7.1.1 (Acyclification of iSCM)**.** *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM with causal graph $G(M)$. For each strongly connected component $C \subseteq V$ of $G(M)$ (i.e., a set of the form $\mathrm{Sc}^{G(M)}(v)$ for $v \in V$), let $g^{[C]} : \mathcal{X}_{J\cup(V\setminus C)\cup W} \to \mathcal{X}_C$ be an (essentially unique) partial solution function of M w.r.t. C. Define $\tilde{f} : \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_W \to \mathcal{X}_V$ by its components*

$$\tilde{f}_v(x_J, x_V, x_W) = g_v^{[\mathrm{Sc}^{G(M)}(v)]}(x_J, x_{V\setminus C}, x_W)$$

*for $v \in V$. $M^{\mathrm{acy}} = (J, V, W, \mathcal{X}, P, \tilde{f})$ is called an **acyclification of** M.*

Note that an acyclification is unique up to iSCM equivalence. The crucial property of this definition is the following result, which also motivates its name.

Proposition 7.1.2. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Its acyclifications M^{acy} are acyclic and observably equivalent to M.*

Proof. Without loss of generality, we may assume that each component f_v only depends on the parents of v:

$$f_v(x) = f_v(x_{\mathrm{Pa}^{G^+(M)}(v)})$$

by means of Remark 6.9.5.

We construct a directed graph S from $G := G(M)$ with its strongly connected components $\{\mathrm{Sc}^G(v) : v \in V \cup J \cup W\}$ as nodes, and directed edges $C \longrightarrow D$ if there is a directed edge $c \longrightarrow d$ in G with $c \in C, d \in D$ and $C \neq D$. The graph S cannot contain a directed cycle, as that would imply the existence of a directed cycle in G that traverses more than one of its strongly connected components. Hence S is a DAG.

Choose a topological ordering $<$ of S. Any node C in S can only have incoming directed edges in S from $\mathrm{Pred}^S_<(C)$. This implies that for $v \in V$, $C = \mathrm{Sc}^G(v)$ can only have incoming edges in G from $\bigcup \mathrm{Pred}^S_\leq(C)$. That implies that the causal mechanism f_C only depends on variables in $\bigcup \mathrm{Pred}^S_\leq(C)$, and hence we can pick a version of the partial solution function g_C which also depends only on variables in $\bigcup \mathrm{Pred}^S_<(C)$. Then $\tilde{f}_C$ only depends on variables in $\bigcup \mathrm{Pred}^S_<(C)$ only. Therefore, for $v, w \in V$, a directed edge $w \longrightarrow v$ in $G(M^{\mathrm{acy}})$ implies $w \in \bigcup \mathrm{Pred}^S_<(\mathrm{Sc}^G(v))$. We can therefore refine the topological ordering $<$ of S to a topological ordering of $G(M^{\mathrm{acy}})$, by arbitrarily ordering the nodes within each strongly connected component of G. Hence $G(M^{\mathrm{acy}})$ is acyclic.

M and M^{acy} are observably equivalent by construction. This follows since the following equivalences hold M-a.s.:

$$
\begin{aligned}
& x = \tilde{f}(x) \\
\iff & \forall C \in S, C \subseteq V : x_C = \tilde{f}_C(x) \\
\iff & \forall C \in S, C \subseteq V : x_C = g^{[C]}(x_J, x_{V \setminus C}, x_W) \\
\iff & \forall C \in S, C \subseteq V : x_C = f_C(x) \\
\iff & x = f(x).
\end{aligned}
$$

□

The iSCM notion of acyclification is compatible with the graphical notion of acyclification:

Proposition 7.1.3. *Let M be a simple iSCM. Then $G^+(M^{\text{acy}}) \subseteq \tilde{G}^+$ for any graphical acyclification $\tilde{G}^+$ of $G^+(M)$, and $G(M^{\text{acy}}) \subseteq \tilde{G}$ for any graphical acyclification $\tilde{G}$ of $G(M)$.*

Proof. Write $G := G(M)$. Let $\tilde{G}^+$ be a graphical acyclification of $G^+(M)$. By definition, $G^+(M^{\text{acy}})$ and $\tilde{G}^+$ have the same nodes (input nodes J and output nodes $V \cup W$). $G^+(M^{\text{acy}})$ has no bidirected edges, but $\tilde{G}^+$ might. If there is a directed edge $i \to j$ in $G^+(M^{\text{acy}})$ with $i \in J \cup V \cup W$ and $j \in V$, then the solution function $g^{[\text{Sc}^G(j)]}$ of $M^{[\text{Sc}^G(j)]}$ essentially depends on x_i. This can only happen if $i \notin \text{Sc}^G(j)$ and i is a parent of some k according to M with $k \in \text{Sc}^G(j)$, i.e., if $i \to k$ in G. In that case, $i \to j$ in $\tilde{G}^+$ by definition of the graphical acyclification.

Let $\tilde{G}$ be a graphical acyclification of $G(M)$. By definition, $G(M^{\text{acy}})$ and $\tilde{G}$ have the same nodes (input nodes J and output nodes V). If $G(M^{\text{acy}})$ has a bidirected edge $i \leftrightarrow j$ (with $i, j \in V$) then there exists a $k \in W$ such that $i \leftarrow k \to j$ in $G^+(M^{\text{acy}})$. Then the solution function $g^{[\text{Sc}^G(i)]}$ of $M^{[\text{Sc}^G(i)]}$ essentially depends on x_k, and the solution function $g^{[\text{Sc}^G(j)]}$ of $M^{[\text{Sc}^G(j)]}$ essentially depends on x_k. This implies that there exist $i' \in \text{Sc}^G(i)$ and $j' \in \text{Sc}^G(j)$ such that k is a parent of both i' and j' according to M, i.e., $i' \leftarrow k \to j'$ in $G^+(M)$. Therefore, the bidirected edge $i' \leftrightarrow j'$ is present in $G(M)$. By definition, then, $i \leftrightarrow j$ must be present in $\tilde{G}$. If there is a directed edge $i \to j$ in $G(M^{\text{acy}})$ with $i \in J \cup V$ and $j \in V$, then the solution function $g^{[\text{Sc}^G(j)]}$ of $M^{[\text{Sc}^G(j)]}$ essentially depends on x_i. This can only happen if $i \notin \text{Sc}^G(j)$ and i is a parent of some j' according to M with $j' \in \text{Sc}^G(j)$, i.e., if $i \to j'$ in $G(M)$. In that case, $i \to j$ in $\tilde{G}$ by definition of the graphical acyclification. □

Hence, two nodes in the same strongly connected component of $G^+(M)$ do not have any edge between them in $G^+(M^{\text{acy}})$, whereas they necessarily have a connecting edge in any acyclification $\tilde{G}^+$ of $G^+(M)$. For two nodes in different strongly connected components of $G^+(M)$, the edges in $G^+(M^{\text{acy}})$ are also present in $\tilde{G}^+$, but not necessarily vice versa, as some parent-relations may cancel out in the solution function.

7.2. Markov Properties

With the help of the acyclifications, we can easily derive a Markov property for simple iSCMs from the Markov property for CBNs by reducing the general cyclic case to an acyclic case.

Corollary 7.2.1 (Global Markov property for simple iSCMs)**.** *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM with graph $G^+(M)$ and Markov kernel $P_M(X_V, X_W \,|\, \mathrm{do}(X_J))$. Then for all $A, B, C \subseteq J \cup V \cup W$ (not necessarily disjoint) we have the implication:*

$$A \overset{\sigma}{\underset{G^+(M)}{\perp}} B \,|\, C \implies X_A \underset{P_M(X_V, X_W \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C.$$

If one wants to make the implicit dependence on J more explicit one can equivalently also write:

$$A \overset{\sigma}{\underset{G^+(M)}{\perp}} J \cup B \,|\, C \implies X_A \underset{P_M(X_V, X_W \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_J, X_B \,|\, X_C.$$

For the causal graph $G(M)$, we get similar statements: for all $A, B, C \subseteq J \cup V$ (not necessarily disjoint),

$$A \overset{\sigma}{\underset{G(M)}{\perp}} B \,|\, C \implies X_A \underset{P_M(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C$$

or, equivalently,[44]

$$A \overset{\sigma}{\underset{G(M)}{\perp}} J \cup B \,|\, C \implies X_A \underset{P_M(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_J, X_B \,|\, X_C.$$

Proof. Choose an acyclification $\tilde{G}^+$ of $G^+(M)$. Then:

$$\begin{aligned}
A \overset{\sigma}{\underset{G^+(M)}{\perp}} B \,|\, C &\iff A \overset{d}{\underset{\tilde{G}^+}{\perp}} B \,|\, C \\
&\implies A \overset{d}{\underset{G^+(M^{\mathrm{acy}})}{\perp}} B \,|\, C \\
&\implies X_A \underset{P_{M^{\mathrm{acy}}}(X_V, X_W \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C \\
&\iff X_A \underset{P_M(X_V, X_W \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C.
\end{aligned}$$

For the various implications / equivalences, we used:

1. $\tilde{G}^+$ is an acyclification of $G^+(M)$ together with Proposition 3.5.2;
2. $G^+(M^{\mathrm{acy}}) \subseteq \tilde{G}^+$ from Proposition 7.1.3, and that removing edges cannot turn a d-separation into a d-connection;

[44] Occassionally, we will use a shorthand notation, writing $\perp\!\!\!\perp_{P_M}$ instead of the longer $\perp\!\!\!\perp_{P_M(X_V \,|\, \mathrm{do}(X_J))}$.

3. the global Markov property Theorem 4.2.1 for M^{acy} interpreted as a causal Bayesian network as in the proof of Proposition 6.11.4 point i) (with deterministic Markov kernels for the endogenous variables, and purely probabilistic Markov kernels for the exogenous random variables), exploiting Proposition 7.1.2 that states that the acyclification M^{acy} is acyclic;

4. by Proposition 7.1.2, the acyclification M^{acy} has the same Markov kernel as the original iSCM M.

The Markov property for $G(M)$ follows as a special case by restricting $A, B, C \subseteq J \cup V$, and noting that

$$A \overset{\sigma}{\underset{G(M)}{\perp}} B \,|\, C \implies A \overset{\sigma}{\underset{G^+(M)}{\perp}} B \,|\, C$$

and that

$$X_A \underset{P_M(X_V, X_W \,|\, \text{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C \implies X_A \underset{P_M(X_V \,|\, \text{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C.$$

□

7.3. Do-calculus for simple iSCMs

With the global Markov property for simple iSCMs, it becomes straightforward to derive the do-calculus for simple iSCMs. First we will introduce some notation. The setting will be that a simple iSCM $M = (J, V, W, \mathcal{X}, P, f)$ is given. For $B \subseteq V$, we will introduce intervention variables $(I_b)_{b \in B}$ to jointly encode different intervened iSCMs $\{M_{\text{do}(X_C)} : C \subseteq B\}$ into a single iSCM $M_{\text{do}(I_B)}$ as in Definition 6.4.13. We will denote the causal graph of M by $G := G(M)$ and the causal graph of $M_{\text{do}(I_B)}$ by $G_{\text{do}(I_B)} = G(M_{\text{do}(I_B)})$. In the do-calculus, we make use of the extended graph $G_{\text{do}(I_B)}$ to test the separation statement, while the conclusion about the properties of certain Markov kernels concerns those of the original iSCM M.

The Markov kernel of simple iSCM M exists, is unique, and is denoted by $P_M(X_V \mid \text{do}(X_J))$. For a subset $T \subseteq V$, we write

$$P_M(X_{V \setminus T} \mid \text{do}(X_J, X_T)) := P_{M_{\text{do}(T)}}(X_{V \setminus T} \mid \text{do}(X_J, X_T)).$$

By conditioning on a subset $S \subseteq V \setminus T$, we obtain a conditional Markov kernel

$$P_M(X_{V \setminus (T \cup S)} \mid X_S, \text{do}(X_J, X_T))$$

which is unique up to a $P_M(X_S \mid \text{do}(X_J, X_T))$-null set.

The only modification to the do-calculus for simple iSCMs as compared to that for causal Bayesian networks is that we have to replace all d-separations by σ-separations.

Theorem 7.3.1 (Almost-sure do-calculus for simple iSCMs, simplified)**.** *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM with causal graph $G = G(M)$. Assume that we have σ-finite reference measures μ_v on $\mathcal{X}_v$ for every $v \in V$ and put $\mu_F := \bigotimes_{v \in F} \mu_v$ for $F \subseteq V$. Let $A, B, C \subseteq V$ and $D \subseteq V \cup J$ be such that A, B, C, D are pairwise disjoint. Then we have the following 4 rules relating Markov kernels that can be generated from the iSCM:*

1. ***Insertion/deletion of observation***, *for* $J \subseteq D$*: if*

$$A \underset{G_{\mathrm{do}(D)}}{\overset{\sigma}{\perp}} B \,|\, C \cup D, \qquad \mu_{B \cup C} \ll P_M(X_B, X_C |\, \mathrm{do}(X_D)) \ll \mu_{B \cup C},$$

then:

$$P_M(X_A | X_B, X_C, \mathrm{do}(X_D)) = P_M(X_A | X_C, \mathrm{do}(X_D)) \qquad \mu_{B \cup C}\text{-a.s.}.$$

2. ***Action/observation exchange***, *for* $J \subseteq D$*: if*

$$A \underset{G_{\mathrm{do}(I_B, D)}}{\overset{\sigma}{\perp}} I_B \,|\, B \cup C \cup D, \qquad \mu_{B \cup C} \ll P_M(X_B, X_C |\, \mathrm{do}(X_D)) \ll \mu_{B \cup C},$$

$$\mu_C \ll P_M(X_C |\, \mathrm{do}(X_B, X_D)) \ll \mu_C,$$

then:

$$P_M(X_A | X_B, X_C, \mathrm{do}(X_D)) = P_M(X_A |\, \mathrm{do}(X_B), X_C, \mathrm{do}(X_D)) \qquad \mu_{B \cup C}\text{-a.s.}.$$

3. ***Insertion/deletion of action***, *for* $J \subseteq D$*: if*

$$A \underset{G_{\mathrm{do}(I_B, D)}}{\overset{\sigma}{\perp}} I_B \,|\, C \cup D, \qquad \mu_C \ll P_M(X_C |\, \mathrm{do}(X_B, X_D)) \ll \mu_C,$$

$$\mu_C \ll P_M(X_C |\, \mathrm{do}(X_D)) \ll \mu_C,$$

then

$$P_M(X_A |\, \mathrm{do}(X_B), X_C, \mathrm{do}(X_D)) = P_M(X_A | X_C, \mathrm{do}(X_D)) \qquad \mu_C\text{-a.s.}.$$

4. ***Deletion of input***: *If*

$$A \underset{G_{\mathrm{do}(D)}}{\overset{\sigma}{\perp}} J \,|\, C \cup D, \qquad \mu_C \ll P_M(X_C |\, \mathrm{do}(X_{D \cup J})) \ll \mu_C,$$

then there exists a Markov kernel $P_M(X_A | X_C, \mathrm{do}(X_D, \cancel{X_{J \setminus D}}))$ *such that:*

$$P_M(X_A | X_C, \mathrm{do}(X_D, \cancel{X_{J \setminus D}})) = P_M(X_A | X_C, \mathrm{do}(X_{D \cup J})) \qquad \mu_C\text{-a.s.}.$$

Proof. The proof is analogous to that of Corollary 5.1.3, except that it applies the global Markov property for simple iSCMs, Corollary 7.2.1, instead of the one for causal Bayesian networks, Theorem 4.2.1. □

While the derivation of the do-calculus relies essentially on the global Markov property, sometimes one can make use of the global Markov property and a more careful analysis of null sets to obtain stronger conclusions. In particular, Proposition 5.1.8 and Theorem 5.1.2 also hold for simple iSCMs if one replaces the d-separation statements by the analogous σ-separation statements (but we will not bother to write out these statements in detail).

7.4. Adjustment

Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let us assume $J \subseteq D$. We are interested in estimating the conditional causal effect:

$$P_M(X_A | X_C, \mathrm{do}(X_B, X_D)),$$

but we only have data from:

$$P_M(X_A, X_B, X_F | X_C, \mathrm{do}(X_D)).$$

This is an example of a *causal domain adaptation problem.* The following (pairwise disjoint) index sets will have the following roles:

$A \subseteq V$: the outcome variables of interest.

$B \subseteq V$: the treatment or intervention variables.

$C \subseteq V$: general conditional (context) variables under which the data was collected.

$J \subseteq D \subseteq V \cup J$: general interventional (context) variables that were set by the experimenter.

$F_0 \subseteq V$: core adjustment variables, i.e. features that were measured.

$F_1 \subseteq V$: additional measured adjustment variables, with $F = F_0 \cup F_1$.

Although we could assume additional unobserved variables $H \subseteq V \cup W$ like for Theorem 5.2.3, the (positivity and independence) assumptions regarding these latent variables are rather strong, so for simplicity and without loss of too much (practical) generality, we will here only consider the special case $H = \emptyset$. We will make use of the same extended iSCM $M_{\mathrm{do}(I_B)}$ with intervention variables I_b for $b \in B$ and graph $G_{\mathrm{do}(I_B)}$ as for stating the do-calculus.

Theorem 7.4.1 (General adjustment formula for simple iSCMs)**.** *Given a simple iSCM $M = (J, V, W, \mathcal{X}, P, f)$ with causal graph $G = G(M)$. Assume that all the following σ-separations hold in the graph $G_{\mathrm{do}(I_B, D)}$:*

$$F_0 \underset{G_{\mathrm{do}(I_B,D)}}{\overset{\sigma}{\perp}} I_B \,|(C \cup D), \tag{43}$$

$$A \underset{G_{\mathrm{do}(I_B,D)}}{\overset{\sigma}{\perp}} (F_1 \cup I_B) \,|(B \cup F_0 \cup C \cup D). \tag{44}$$

Further assume that we have reference measures μ_v on $\mathcal{X}_v$, $v \in V$, such that:

$$\mu_{B \cup C \cup F} \ll P(X_B, X_C, X_F | \,\mathrm{do}(X_D)) \ll \mu_{B \cup C \cup F},$$
$$\mu_{C \cup F} \ll P(X_C, X_F | \,\mathrm{do}(X_B, X_D)) \ll \mu_{C \cup F}.$$

Then we have the adjustment formula:

$$P_M(X_A | X_C, \mathrm{do}(X_B, X_D)) = P_M(X_A | X_B, X_C, X_F, \mathrm{do}(X_D)) \circ P_M(X_F | X_C, \mathrm{do}(X_D)) \qquad \mu_{B \cup C}\text{-a.s.}$$

Proof. Analogous to that of Theorem 5.2.3, but now using the global Markov property for simple iSCMs, Corollary 7.2.1, instead of the one for causal Bayesian networks, Theorem 4.2.1. □

For the special case $F_1 = C = \emptyset$, by a direct more careful analysis we get a version with weaker positivity assumptions:

Theorem 7.4.2 (Interventional backdoor covariate adjustment formula)**.** *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM with causal graph $G = G(M)$. Assume that the* ***interventional backdoor criterion*** *in the graph $G_{\mathrm{do}(I_B,D)}$ holds:*

1. $F \underset{G_{\mathrm{do}(I_B,D)}}{\overset{\sigma}{\perp}} I_B \,|\, D$, *and:*

2. $A \underset{G_{\mathrm{do}(I_B,D)}}{\overset{\sigma}{\perp}} I_B \,|(B \cup F \cup D)$.

Further assume the following absolute continuity:

$$P_M(X_F|\,\mathrm{do}(X_D)) \otimes P_M(X_B|\,\mathrm{do}(X_D)) \ll P_M(X_F, X_B|\,\mathrm{do}(X_D)).$$

Then we have the adjustment formulas:

$$P_M(X_A, X_F|\,\mathrm{do}(X_B, X_D)) = P_M(X_A|X_F, X_B, \mathrm{do}(X_D)) \otimes P_M(X_F|\,\mathrm{do}(X_D)) \quad P_M(X_B|\,\mathrm{do}(X_D))\text{-a.s.},$$
$$P_M(X_A|\,\mathrm{do}(X_B, X_D)) = P_M(X_A|X_F, X_B, \mathrm{do}(X_D)) \circ P_M(X_F|\,\mathrm{do}(X_D)) \quad P_M(X_B|\,\mathrm{do}(X_D))\text{-a.s.}$$

Proof. Analogous to that of Corollary 5.2.5, but now using the global Markov property for simple iSCMs, Corollary 7.2.1, instead of the one for causal Bayesian networks, Theorem 4.2.1. □

We can now further specialize to the case with $F_1 = C = D = J = \emptyset$ and immediately get:

Corollary 7.4.3 (Backdoor covariate adjustment for simple iSCMs)**.** *Given a simple iSCM $M = (J, V, W, \mathcal{X}, P, f)$ with causal graph $G = G(M)$. Assume that the* ***backdoor criterion*** *holds:*

1. $F \underset{G_{\mathrm{do}(I_B)}}{\overset{\sigma}{\perp}} I_B$, *and:*

2. $A \underset{G_{\mathrm{do}(I_B)}}{\overset{\sigma}{\perp}} I_B \,|(B \cup F)$.

Further assume the following absolute continuity:

$$P_M(X_F) \otimes P_M(X_B) \ll P_M(X_F, X_B).$$

Then we have the adjustment formulae:

$$P_M(X_A, X_F|\,\mathrm{do}(X_B)) = P_M(X_A|X_F, X_B) \otimes P_M(X_F) \qquad P_M(X_B)\text{-a.s.},$$
$$P_M(X_A|\,\mathrm{do}(X_B)) = P_M(X_A|X_F, X_B) \circ P_M(X_F) \qquad P_M(X_B)\text{-a.s.}$$

The literature often fails to mention the strict positivity assumptions, even though without sufficient positivity, the various backdoor criteria may not hold. A simple example of how the adjustment formula may fail if the strict positivity assumptions are not met is provided in Example 5.3.29.

7.5. Bounds on causal effects

If causal effects cannot be identified from the observable distribution, we may still be able to derive informative bounds (see also Figure 22). We first prove the *consistency* property of solution functions of simple iSCMs.

Proposition 7.5.1. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ be a (essentially unique) solution function of M. Let $V = A \dot{\cup} B$ be a partition of the endogenous variables of M, and let $g^{[B]} : \mathcal{X}_A \times \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_B$ be a (essentially unique) solution function of M w.r.t. B. Then:*

$$g_B(x_J, x_W) = g^{[B]}(g_A(x_J, x_W), x_J, x_W) \qquad M\text{-a.s.}.$$

Proof. M-a.s.:

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

Also, M-a.s.:

$$x_B = f_B(x) \iff x_B = g^{[B]}(x_A, x_J, x_W).$$

Hence, M-a.s.:

$$\begin{cases} x_A &= f_A(x), \\ x_B &= f_B(x) \end{cases} \implies \begin{cases} x_A &= g_A(x_J, x_W), \\ x_B &= g^{[B]}(x_A, x_J, x_W) \end{cases} \implies \begin{cases} x_A &= g_A(x_J, x_W), \\ x_B &= g^{[B]}(g_A(x_J, x_W), x_J, x_W). \end{cases}$$

The essential uniqueness of the solution function g of M now implies the consistency statement. □

This proposition shows that the "consistency assumption" commonly made in the potential outcomes framework holds true for potential outcomes induced by simple iSCMs.

[MN98] proved the following 'natural' bounds. We point out here that they also hold for simple iSCMs.

Theorem 7.5.2 (Natural bounds on causal effect)**.** *Let M be a simple iSCM with endogenous variables $V \supseteq \{1, 2, 3\}$ and no exogenous input variables. Assume that $\mathcal{X}_1$ is discrete. Then*

$$\begin{aligned} P_M(X_1 = a, X_2 \in B | X_3 \in C) &\leq P_M(X_2 \in B | X_3 \in C, \mathrm{do}(X_1 = a)) \\ &\leq P_M(X_1 = a, X_2 \in B | X_3 \in C) + P_M(X_1 \neq a | X_3 \in C). \end{aligned} \tag{45}$$

for any $a \in \mathcal{X}_1$, measurable $B \subseteq \mathcal{X}_2$, and measurable $C \subseteq \mathcal{X}_3$ with $P_M(X_3 \in C) > 0$.

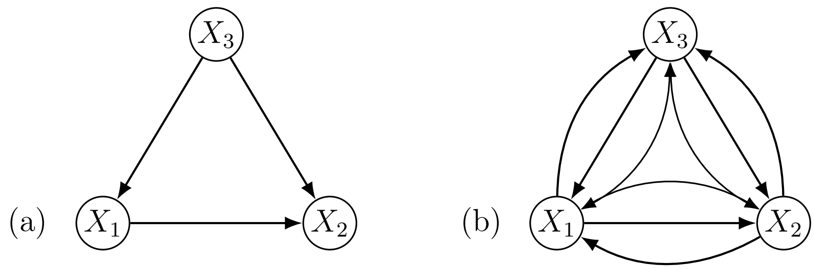


Figure 22: Two causal graphs of simple iSCMs. In (a), the Markov kernel $P_M(X_2 \,|\, \mathrm{do}(X_1), X_3)$ is identifiable (under positivity assumptions) from $P_M(X_1, X_2, X_3)$. In (b), it is not identifiable, but we can still bound it using the natural bounds if both X_1 and X_3 are discrete.

Proof. We use the consistency $X_2 = X_2^{\mathrm{do}(X_1=X_1)}$ and elementary probability theory:

$$
\begin{aligned}
P_M(X_1 = a, X_2 \in B | X_3 \in C) &= P_M(X_1 = a, X_2^{\mathrm{do}(X_1=a)} \in B | X_3 \in C) \\
&\leq P_M(X_2^{\mathrm{do}(X_1=a)} \in B | X_3 \in C) \\
&= P_M(X_1 = a, X_2^{\mathrm{do}(X_1=a)} \in B | X_3 \in C) \\
&\quad + P_M(X_1 \neq a, X_2^{\mathrm{do}(X_2=a)} \in B | X_3 \in C) \\
&\leq P_M(X_1 = a, X_2 \in B | X_3 \in C) + P_M(X_1 \neq a | X_3 \in C)
\end{aligned}
$$

The statement follows since

$$
P_M(X_2 \in B | X_3 \in C, \mathrm{do}(X_1 = a)) = P_M(X_2^{\mathrm{do}(X_1=a)} \in B | X_3 \in C).
$$

□

This so-called "natural" bound can be shown to be tight. Remarkably, we do not need to make any assumptions regarding the causal relations between the three endogenous variables. This allows us to bound the causal effect of X_1 on X_2 in the presence of confounding and cycles. Unfortunately, it can be shown that there exists no analogous bound in case X_1 is real-valued.

If one has a priori knowledge about the range of X_2 (in case it is real-valued), one can also derive a bound on the expected result of an intervention [MP13].

Corollary 7.5.3. *In the situation of Theorem 7.5.2, suppose that $\mathcal{X}_2 = [\alpha, \beta]$ and $0 < P_M(X_1 = a | X_3 \in C) < 1$. Then*

$$
\begin{aligned}
&P_M(X_1 = a | X_3 \in C)\mathbb{E}_M(X_2 | X_1 = a, X_3 \in C) + \alpha P_M(X_1 \neq a | X_3 \in C) \\
\leq\ &\mathbb{E}_M(X_2 | X_3 \in C, \mathrm{do}(X_1 = a)) \\
\leq\ &P_M(X_1 = a | X_3 \in C)\mathbb{E}_M(X_2 | X_1 = a, X_3 \in C) + \beta P_M(X_1 \neq a | X_3 \in C).
\end{aligned} \tag{46}
$$

Proof. Using consistency, we get:

$$
\begin{aligned}
&P_M(X_2^{\mathrm{do}(X_1=a)} \in B|X_3 \in C) \\
&= P_M(X_1 = a|X_3 \in C)P_M(X_2^{\mathrm{do}(X_1=a)} \in B|X_1 = a, X_3 \in C) \\
&+ P_M(X_1 \neq a|X_3 \in C)P_M(X_2^{\mathrm{do}(X_1=a)} \in B|X_1 \neq a, X_3 \in C) \\
&= P_M(X_1 = a|X_3 \in C)P_M(X_2 \in B|X_1 = a, X_3 \in C) \\
&+ P_M(X_1 \neq a|X_3 \in C)P_M(X_2^{\mathrm{do}(X_1=a)} \in B|X_1 \neq a, X_3 \in C).
\end{aligned}
$$

Integrating over X_2, the assumption $\mathcal{X}_2 = [\alpha, \beta]$, interval arithmetic, and affinity of expected values, gives:

$$
\begin{aligned}
&\mathbb{E}_M(X_2^{\mathrm{do}(X_1=a)}|X_3 \in C) \\
&= P_M(X_1 = a|X_3 \in C)\mathbb{E}_M(X_2|X_1 = a, X_3 \in C) \\
&+ P_M(X_1 \neq a|X_3 \in C)\mathbb{E}_M(X_2^{\mathrm{do}(X_1=a)}|X_1 \neq a, X_3 \in C) \\
&\in P_M(X_1 = a|X_3 \in C)\mathbb{E}_M(X_2|X_1 = a, X_3 \in C) + P_M(X_1 \neq a|X_3 \in C)[\alpha, \beta].
\end{aligned}
$$

□

8. Counterfactuals with iSCMs

Counterfactuals are questions of the kind "Fred obtained a cum laude for his PhD; would he have obtained the distinction also if he were female?". Counterfactuals consider a hypothetical situation that is "contrary to the fact", that is, which differs from what was actually observed. One of the big practical obstacles of dealing with counterfactual probabilities is that they are typically not identifiable from experimental data, and at best only bounds on such quantities can be obtained. For systems with (almost) deterministic causal relations, these bounds may become quite informative, but tend to become more loose as the stochasticity in the system increases. While it would thus perhaps be easiest to avoid counterfactuals altogether, they do appear naturally in law and engineering. Humans also have a tendency to communicate using counterfactuals, and the grammar of many languages distinguishes counterfactual statements. This might be an inductive bias towards dealing with (almost) deterministic systems.

In this chapter, we will focus on counterfactuals that are hypothetical statements (or questions) regarding the effects of some action that is contrary-to-fact, closely following Pearl's approach to counterfactuals. One can consider other types of counterfactuals as well, for example "backtracking" counterfactuals. Dealing with counterfactuals appears to be one of the least well-defined, but perhaps also most intriguing, aspects of causality.

8.1. Modeling counterfactuals via twinning

For example, suppose you are healthy but drank too much beer last night and now suffer from a hangover. A counterfactual statement is then: "If I had not drunk so much beer yesterday, I would feel much better now." This statement invites one to imagine an alternative world in which everything is the same as in the actual world, with the sole difference that you did not drink beer last night. We can then use our causal model of the world to predict the consequences of this action (e.g., since you were in a healthy state and did not drink so much beer, you most likely will feel well in this alternative world).[45] This example already shows the ambiguity typically encountered in counterfactuals: if for you, not drinking beer means that you drink wine instead, then you may actually feel worse than if you had drunk beer.

Indeed, the truth value of such statements is often hard to determine in case the "world" is partially latent or not fully understood. When debugging a computer program, one makes heavy use of counterfactuals: "if I had put a minus sign there, then the output of my program would have been correct". In case the full source code is available, it is in principle straightforward to work out whether such a statement is correct or not, but it becomes more difficult if the full source code is not available. It becomes even more challenging when the output of the computer program is stochastic or it

[45] The word "counterfactual" is also commonly used in the causal inference literature in a weaker sense, describing the potential outcome of a hypothetical situation that is not necessarily "contrary to the fact". For example, some would refer to "If I drink too much beer tonight, I will have a hangover tomorrow" as a counterfactual statement as well. It is important to be aware of this to avoid possible confusion. We will only use the word "counterfactual" in its strong (contrary-to-fact) sense.

is impossible to reproduce its complete input. Generally, in situations where the full causal mechanism is unknown, or exogenous randomness is latent, counterfactuals may not be well-defined quantities. Nevertheless, counterfactual thinking is very common in humans, and toddlers already bombard their parents with counterfactual questions, presumably as a means for them to build internal causal models of the world.

The mathematical formalization of counterfactuals proposed by Pearl provides some clarification, but it also points out their inherent complexity and strong dependence on the chosen model.[46] It mimics the reasoning steps we mentally perform when thinking about counterfactuals by constructing a "(f)actual" world and a parallel "counterfactual" world, which is minimally different in some aspect. The crucial (and often untestable) assumption is that the exogenous random variables have the same *values* in both worlds. A good analogy here is that of two identical twins that share the same latent genetics. Before we give the definition, we will introduce some bookkeeping notation.

Notation 8.1.1. *Given an index set Z we define a primed copy $Z' := \{z' : z \in Z\}$, where each z' is a "primed" copy of z (distinguishable from z itself because of the attached prime symbol). We will also write $(z')^\circ = z$ for $z \in Z$, where the superscript $\circ$ removes the prime, i.e., it maps back to the original of the primed index.*

The following operation on iSCMs (also known as the "twin-network approach" of [BP94b]) provides one possible way of modeling counterfactuals.[47]

Definition 8.1.2. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. We define the* ***twinned iSCM of*** *M as the iSCM $M^{\text{twin}} = \left(J^{\text{twin}} = J \dot\cup J', V^{\text{twin}} = V \dot\cup V', W, \mathcal{X}^{\text{twin}}, P, f^{\text{twin}}\right)$ with $J' = \{j' : j \in J\}$ a copy of J and $V' = \{v' : v \in V\}$ a copy of V, the twinned domain given by*

$$\mathcal{X}^{\text{twin}} = \mathcal{X}_J \times \mathcal{X}_{J'} \times \mathcal{X}_V \times \mathcal{X}_{V'} \times \mathcal{X}_W$$

where $\mathcal{X}_{j'} = \mathcal{X}_j$ for all $j \in J$ and $\mathcal{X}_{v'} = \mathcal{X}_v$ for all $v \in V$, and the twinned causal mechanism components given by

$$f_u^{\text{twin}}\left((x_J, x_{J'}), (x_V, x_{V'}), x_W\right) = \begin{cases} f_u(x_J, x_V, x_W) & u \in V, \\ f_{u^\circ}(x_{J'}, x_{V'}, x_W) & u \in V'. \end{cases}$$

The twinning operation preserves equivalence: $M \equiv \tilde{M} \implies M^{\text{twin}} \equiv \tilde{M}^{\text{twin}}$.

The twinning operation is used to create copies of variables (so that in addition to the one in the factual world, we have its copy in the counterfactual world) that can have different values to describe contrary-to-fact situations. A specific choice in modeling counterfactuals in this way is the assumption that *all exogenous random variables* have

[46] Consider this a warning before attempting to predict counterfactual statements in a data-driven way, for example, using a neural network.

[47] The twinning operation can be applied to any iSCM, but not to any L-CBN, as L-CBNs typically do not explicitly model latent random variables. Only for the subclass of L-CBNs that are in iSCM form (see Definition 4.4.3), i.e., for which every node with at least one parent comes with a deterministic Markov kernel, could we define a twinning operation that is analogous to the one we define for iSCMs.

the same value in the actual and in the counterfactual world. This is a very strong (and typically untestable) assumption.

The English language has a special grammatical construct to express counterfactuals: "If I had studied better, I would have passed the exam," instead of "If I study better, I will pass the exam." For the first statement, we first twin the iSCM and then intervene on it, for the second, we just intervene on the iSCM and there is no need for twinning.[48]

Lemma 8.1.3. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and M^{twin} its twin iSCM.*

1. *Let $\tilde{\pi} : \mathcal{X}_J \times \mathcal{X}_W \to \{0, 1\}$ be a binary property. If $\pi(x_J, x_W)$ holds M-a.s., then $\pi(x_J, x_W)$ holds M^{twin}-a.s. and $\pi(x_{J'}, x_W)$ holds M^{twin}-a.s..*

2. *Let $\pi : \mathcal{X}_J \times \mathcal{X}_W \times \mathcal{X}_V \to \{0, 1\}$ be a binary property. If $\pi(x_J, x_W, x_V)$ holds M-a.s., then $\pi(x_J, x_W, x_V)$ holds M^{twin}-a.s., and $\pi(x_{J'}, x_W, x_{V'})$ holds M^{twin}-a.s..*

Proof. We show the second claim (the first then follows as a special case).

If π holds M-a.s.,

$$A := \{(x_J, x_W, x_V) \in \mathcal{X}_J \times \mathcal{X}_W \times \mathcal{X}_V : \pi(x_J, x_W, x_V) = 0\}$$

is contained in a M-null set of the form $N := \tilde{N} \times \mathcal{X}_V$ with $\tilde{N} \subseteq \mathcal{X}_J \times \mathcal{X}_W$ such that for each $x_J \in \mathcal{X}_J$, the section $\tilde{N}_{x_J} = \{x_W \in \mathcal{X}_W : (x_J, x_W) \in \tilde{N}\}$ is a $P_M(X_W)$-null set.

Note that

$$\{(x_J, x_{J'}, x_W, x_V, x_{V'}) \in \mathcal{X}_J \times \mathcal{X}_{J'} \times \mathcal{X}_W \times \mathcal{X}_V \times \mathcal{X}_{V'} : \pi(x_J, x_W, x_V) = 0\} = \mathcal{X}_{J'} \times A \times \mathcal{X}_{V'}.$$

For this to be a M^{twin}-null set, it must be contained in a M^{twin}-null set of the form $N^{\text{twin}} := \tilde{N}^{\text{twin}} \times \mathcal{X}_V \times \mathcal{X}_{V'}$ with $\tilde{N}^{\text{twin}} \subseteq \mathcal{X}_J \times \mathcal{X}_{J'} \times \mathcal{X}_W$ such that for each $x_J \in \mathcal{X}_J, x_{J'} \in \mathcal{X}_{J'}$, the section $(\tilde{N}^{\text{twin}})_{(x_J, x_{J'})} = \{x_W \in \mathcal{X}_W : (x_J, x_{J'}, x_W) \in \tilde{N}^{\text{twin}}\}$ is a $P_M(X_W)$-null set. Note that $\tilde{N}^{\text{twin}} := \mathcal{X}_{J'} \times \tilde{N}$ does the job, because $(\tilde{N}^{\text{twin}})_{(x_J, x_{J'})} = (\tilde{N})_{x_J}$.

Similarly, since also

$$\{(x_{J'}, x_W, x_{V'}) \in \mathcal{X}_{J'} \times \mathcal{X}_W \times \mathcal{X}_{V'} : \pi(x_{J'}, x_W, x_{V'}) = 0\}$$

is contained in a M-null set,

$$\{(x_J, x_{J'}, x_W, x_V, x_{V'}) \in \mathcal{X}_J \times \mathcal{X}_{J'} \times \mathcal{X}_W \times \mathcal{X}_V \times \mathcal{X}_{V'} : \pi(x_{J'}, x_W, x_{V'}) = 0\}$$

is contained in a M^{twin}-null set. □

[48]Note that in general, when considering two potential outcomes $X^{\text{do}(x_J)}, X^{\text{do}(x'_J)}$ of iSCM M for different inputs x_J, x'_J one does *not* necessarily assume that $X_W^{\text{do}(x_J)} = X_W^{\text{do}(x'_J)}$ (although one does assume that they have the same distribution, that is, $X_W^{\text{do}(x_J)} \sim X_W^{\text{do}(x'_J)}$). It is more transparent to explicitly introduce such counterfactuals with the twinning construction, because this enforces one to think about which variables are shared across potential worlds and which get an (independently resampled, or reevaluated) copy, and avoids implicitly introducing "cross-world" assumptions, that is, assumptions on the joint distribution $P(X^{\text{do}(x_J)}, X^{\text{do}(x'_J)})$.

Lemma 8.1.4. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. Let $L_1 \subseteq V$ and $L_2 \subseteq V$. Write $K_1 := V \setminus L_1$ and $K_2 := V \setminus L_2$. If $g^{[L_1]} : \mathcal{X}_{J \cup K_1} \times \mathcal{X}_W \to \mathcal{X}_{L_1}$ is a solution function of M w.r.t. L_1 and $g^{[L_2]} : \mathcal{X}_{J \cup K_2} \times \mathcal{X}_W \to \mathcal{X}_{L_2}$ is a solution function of M w.r.t. L_2 then*

$$\begin{aligned} g^{\text{twin}} &: \mathcal{X}_{(J \cup K_1) \cup (J' \cup K_2')} \times \mathcal{X}_W \to \mathcal{X}_{L_1} \times \mathcal{X}_{L_2'} \\ &: (x_{(J \cup K_1) \cup (J' \cup K_2')}, x_W) \mapsto \big(g^{[L_1]}(x_{J \cup K_1}, x_W), g^{[L_2]}(x_{J' \cup K_2'}, x_W) \big) \end{aligned}$$

is a solution function of M^{twin} w.r.t. $L_1 \cup L_2'$. If M is essentially uniquely solvable w.r.t. L_1 and also w.r.t. L_2, then M^{twin} is essentially uniquely solvable w.r.t. $L_1 \cup L_2'$.

Proof. Let f^{twin} denote the causal mechanism of M^{twin}. Then, using Lemma 8.1.3, M^{twin}-a.s.:

$$\begin{aligned} & x_{L_1 \cup L_2'} = f^{\text{twin}}_{L_1 \cup L_2'}(x) \\ \iff & \begin{cases} x_{L_1} &= f_{L_1}(x_J, x_V, x_W) \\ x_{L_2'} &= f_{L_2}(x_{J'}, x_{V'}, x_W) \end{cases} \\ \impliedby & \begin{cases} x_{L_1} &= g^{[L_1]}(x_{J \cup K_1}, x_W) \\ x_{L_2'} &= g^{[L_2]}(x_{J' \cup K_2'}, x_W) \end{cases} \\ \iff & \ x_{L_1 \cup L_2'} = g^{\text{twin}}(x_{J \cup K_1}, x_{J' \cup K_2'}, x_W) \end{aligned}$$

For the claim regarding essentially unique solvability, the " $\impliedby$ " becomes an " $\iff$ ". $\square$

The twinning operation is compatible with marginalization.

Proposition 8.1.5. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. For $L \subseteq V$ such that M is essentially uniquely solvable w.r.t. L,*

$$(M_{\setminus L})^{\text{twin}} \equiv (M^{\text{twin}})_{\setminus (L \cup L')}.$$

Proof. This follows by writing out the definitions. Lemma 8.1.4 with $L_1 = L_2 = L$ shows that if M is essentially uniquely solvable w.r.t. L with partial solution function $g^{[L]}$, then M^{twin} is essentially uniquely solvable w.r.t. $L \cup L'$ with partial solution function $g^{[L \cup L']} : \mathcal{X}_J \times \mathcal{X}_{J'} \times \mathcal{X}_W \times \mathcal{X}_{V \setminus L} \times \mathcal{X}_{V' \setminus L'}$ defined by:

$$g^{[L \cup L']}(x_J, x_{J'}, x_W, x_{V \setminus L}, x_{V' \setminus L'}) = \big(g^{[L]}(x_J, x_W, x_{V \setminus L}), g^{[L]}(x_{J'}, x_W, x_{V' \setminus L'}) \big).$$

$\square$

Another important property of the twinning operation is that it preserves unique solvability and simplicity.

Corollary 8.1.6. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. If M is essentially uniquely solvable, then M^{twin} is essentially uniquely solvable. If M is simple, then M^{twin} is simple.*

Several types of hard interventions are compatible with the twinning operation, in the following sense:

Proposition 8.1.7. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM.*

- *For $T \subseteq J \cup V$, $\xi_T \in \mathcal{X}_T$:*

$$(M^{\text{twin}})_{\text{do}(X_T=\xi_T, X_{T'}=\xi_T)} = (M_{\text{do}(X_T=\xi_T)})^{\text{twin}};$$

- *For $T \subseteq J \cup V$:*

$$(M^{\text{twin}})_{\text{do}(T,T')} = (M_{\text{do}(T)})^{\text{twin}};$$

- *For $T \subseteq J \cup V$:*

$$(M^{\text{twin}})_{\text{do}(I_T, I_{T'})} = (M_{\text{do}(I_T)})^{\text{twin}}$$

  (where we identify I'_t with $I_{t'}$ for all $t \in T$).

Proof. These properties follow by writing out the definitions. $\square$

We can also define a twinning operation on graphs. We will only define this for graphs without bidirected edges (limiting this to a graphical version of the twinning that can be applied to the graph $G^+(M)$ of an iSCM M).

Definition 8.1.8. *Let $G = (J, V \cup W, E)$ be a CDG such that $\text{Pa}^G(W) = \emptyset$. Write $J' := \{j' : j \in J\}$ and $V' := \{v' : v \in V\}$ for copies of J and V, respectively. The **twinned graph** $G^{\text{twin}(J,V)}$ is defined as the CDG $\left(J \dot{\cup} J', V \dot{\cup} V' \cup W, E^{\text{twin}}\right)$ with directed edges*

$$E^{\text{twin}} := E \cup \{w \longrightarrow v' : w \in W, v \in V, w \longrightarrow v \in E\} \cup \{i' \longrightarrow v' : i \in J \cup V, v \in V, i \longrightarrow v \in E\}.$$

In words, we copy the nodes $J \cup V$ (but not the nodes W) and copy the edges accordingly.

The graphical and the iSCM twinning operations are compatible:

Proposition 8.1.9. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. Then*

$$(G^+(M))^{\text{twin}(J \cup V)} = G^+(M^{\text{twin}}).$$

Proof. This follows by writing out the definitions. $\square$

The simplest non-trivial example of a twin iSCM is the following.

Example 8.1.10. *Consider an iSCM with exogenous input variable X, endogenous variable Y, exogenous random variable W, and structural equation:*

$$Y^{\text{do}(x)} = f(x, W)$$

Its graph is depicted in Figure 23. Twinning adds an exogenous input variable X', an endogenous variable Y', and structural equation

$$(Y')^{\text{do}(x')} = f(x', W)$$

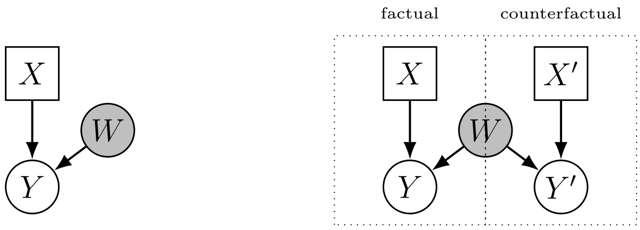


Figure 23: Left: Graph $G^+(M)$ of iSCM M. Right: Graph $G^+(M^{\text{twin}}) = (G^+(M))^{\text{twin}(X,Y)}$ of the twinned iSCM M^{twin}. The exogenous random variable W is shared between the factual and counterfactual "world", whereas the exogenous input variable X and the endogenous (output) variable Y may differ between the two worlds (with X' and Y' denoting the corresponding variables in the counterfactual "world").

but keeps the same endogenous random variable W. The graph of the twinned iSCM is also shown in Figure 23.

For instance, X could be the number of glasses of beer you consumed yesterday evening, Y the severity of a headache the next morning, and W would represent all other possible causes of a headache (e.g., COVID-19, a concussion obtained in a rugby game, the number of glasses of wine you consumed yesterday evening, . . .). When stating "If I had not drunk so much beer yesterday, I would feel much better now," we imagine a counterfactual world in which the value of W is the same as in the (f)actual world, but X' (and therefore Y') may have different values in the counterfactual world than the corresponding values X (and Y) in the factual world.

We are not claiming that *all* counterfactuals can be modeled naturally using a twinned iSCM. For example, the difference between the actual and the counterfactual world may not always be due to an intervention, that overrides certain causal mechanisms. Sometimes, the causal mechanisms in both worlds remain the same, but they differ with respect to certain initial conditions or background conditions.

Example 8.1.11. *The counterfactual: "Yesterday I met an old friend in the street that I hadn't seen in years. We decided to go have a beer in the pub, even though I had plans to study for my exam that night. If I had not bumped into my old friend, I would not have ended up in the pub yesterday evening." is an example of a **backtracking** counterfactual. The aspect that is different in the coutnerfactual world is that by chance, I met my old friend. It seems unnatural to model this as an action.*

Backtracking counterfactuals are actively studied by philosophers and logicians. We will not attempt to formalize them here.

8.2. Counterfactual Equivalence

In Definition 6.7.4, we defined the notions of observable and interventional equivalence for simple iSCMs. We can add a more fine-grained notion of equivalence by making use of the twinning operation, which we refer to as counterfactual equivalence.[49]

Definition 8.2.1. *Let $M = (J, V, W, \mathcal{X}, P, f)$ and $\tilde{M} = \left(\tilde{J}, \tilde{V}, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f}\right)$ be two simple iSCMs and $O \subseteq V \cap \tilde{V}$ a subset. We say that M and $\tilde{M}$ are* ***counterfactually equivalent w.r.t.*** *O if the twin iSCMs M^{twin} and $\tilde{M}^{\text{twin}}$ are interventionally equivalent w.r.t. $O \cup O'$, where $O' \subseteq V' \cap \tilde{V}'$ is the copy of O.*

More generally, one could define counterfactual equivalence not only with respect to an observed set of variables, but also with respect to a given set of interventions.

We get the following important corollary of Theorem 6.8.10.

Corollary 8.2.2. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM, $L \subseteq V$, and $M_{\setminus L}$ its marginalization over L. Then M and $M_{\setminus L}$ are observably, interventionally and counterfactually equivalent w.r.t. $V \setminus L$.*

Proof. The observable and interventional equivalence is the claim of Theorem 6.8.10. Write $K = V \setminus L$. By Proposition 8.1.5, $(M_{\setminus L})^{\text{twin}} \equiv (M^{\text{twin}})_{\setminus(L \cup L')}$, and hence the two are interventionally equivalent w.r.t. $K \cup K'$. Since M^{twin} and its marginalization $(M^{\text{twin}})_{\setminus(L \cup L')}$ are interventionally equivalent w.r.t. $K \cup K'$ by Theorem 6.8.10, M and $M_{\setminus L}$ are counterfactually equivalent w.r.t. K. □

Lemma 8.2.3. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Then M and M^{twin} are interventionally equivalent w.r.t. V.*

Proof. We have to show that for any $T \subseteq V$, $M_{\text{do}(T)}$ and $(M^{\text{twin}})_{\text{do}(T)}$ are observably equivalent w.r.t. $V \setminus T$. Let $T \subseteq V$. Let $g_{\text{do}(T)} : \mathcal{X}_{J \cup T} \times \mathcal{X}_W \to \mathcal{X}_{V \setminus T}$ be an (essentially unique) solution function of M w.r.t. $V \setminus T$ and $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ the (essentially unique) solution function of M.

By Lemma 8.1.4 with $L_1 = V \setminus T$, $L_2 = V$,

$$\begin{aligned}\tilde{g} &: \mathcal{X}_{(J \cup T) \cup J'} \times \mathcal{X}_W \to \mathcal{X}_{V \setminus T} \times \mathcal{X}_{V'} \\ &: (x_{(J \cup T) \cup J'}, x_W) \mapsto \left(g_{\text{do}(T)}(x_{J \cup T}, x_W), g(x_{J'}, x_W)\right)\end{aligned}$$

is the (essentially) unique solution function of (M^{twin}) w.r.t. $(V \setminus T) \cup V'$. Note that $g_{\text{do}(T)} \circ \text{pr}_{\mathcal{X}_{J \cup T} \times \mathcal{X}_W} = \tilde{g}_{V \setminus T}$.

Then $P_{M_{\text{do}(T)}}(X_{V \setminus T} \mid \text{do}(X_{J \cup T}))$ is the pushforward $(g_{\text{do}(T)})_*(P)$ of the exogenous distribution of $M_{\text{do}(T)}$ (interpreted as a constant Markov kernel $\mathcal{X}_{J \cup T} \dashrightarrow \mathcal{X}_W$). We obtain the Markov kernel $P_{(M^{\text{twin}})_{\text{do}(T)}}(X_{V \setminus T} \mid \text{do}(X_{J \cup T \cup J'}))$ as the pushforward $(\tilde{g}_{V \setminus T})_*(P)$ of the exogenous distribution P of $(M^{\text{twin}})_{\text{do}(T)}$, now interpreted as a constant Markov kernel $\mathcal{X}_{J \cup T} \times \mathcal{X}_{J'} \to \mathcal{X}_W$. Since $g_{\text{do}(T)} \circ \text{pr}_{\mathcal{X}_{J \cup T} \times \mathcal{X}_W} = \tilde{g}_{V \setminus T}$, we obtain the desired conclusion. □

Equivalence is a stronger equivalence relation than counterfactual equivalence, and counterfactual equivalence is a stronger equivalence relation than interventional equivalence.

[49] The definition of counterfactual equivalence for (possibly non-simple) iSCMs is provided in [BFPM21] for iSCMs without exogenous input variables.

Proposition 8.2.4. *For simple iSCMs $M, \tilde{M}$ and a subset $O \subseteq V \cap \tilde{V}$:*

1. *If $M \equiv \tilde{M}$ then M and $\tilde{M}$ are counterfactually equivalent w.r.t. O.*
2. *If M and $\tilde{M}$ are counterfactually equivalent w.r.t. O then M and $\tilde{M}$ are interventionally equivalent w.r.t. O.*

Proof. 1. Suppose $M \equiv \tilde{M}$. Then $M^{\text{twin}} \equiv \tilde{M}^{\text{twin}}$, and the claim directly follows from Proposition 6.7.6.

2. Let $O' \in V' \cap \tilde{V}'$ be the copy of O. Suppose M and $\tilde{M}$ are counterfactually equivalent w.r.t. O. Then M^{twin} and $\tilde{M}^{\text{twin}}$ are interventionally equivalent w.r.t. $O \cup O'$. For every $T \subseteq O \cup O'$, $(M^{\text{twin}})_{\text{do}(T)}$ and $(\tilde{M}^{\text{twin}})_{\text{do}(T)}$ are observably equivalent w.r.t. $(O \cup O') \setminus T$.

   We have to show that M and $\tilde{M}$ are interventionally equivalent w.r.t. O. That is, we have to show that for every $S \subseteq O$, $M_{\text{do}(S)}$ and $\tilde{M}_{\text{do}(S)}$ are observably equivalent w.r.t. $O \setminus S$.

   Now take $S \subseteq O$. Then (using Proposition 8.1.7),

$$(M_{\text{do}(S)})^{\text{twin}} = (M^{\text{twin}})_{\text{do}(S,S')}$$

   and

$$(\tilde{M}_{\text{do}(S)})^{\text{twin}} = (\tilde{M}^{\text{twin}})_{\text{do}(S,S')}$$

   are observably equivalent w.r.t. $(O \cup O') \setminus (S \cup S') = (O \setminus S) \cup (O' \setminus S')$, and hence w.r.t. $O \setminus S$. By Lemma 8.2.3, $(M_{\text{do}(S)})^{\text{twin}}$ and $M_{\text{do}(S)}$ are interventionally equivalent w.r.t. $V \setminus S$. Hence $(M_{\text{do}(S)})^{\text{twin}}$ and $M_{\text{do}(S)}$ are observably equivalent w.r.t. $O \setminus S$. Similarly, $(\tilde{M}_{\text{do}(S)})^{\text{twin}}$ and $\tilde{M}_{\text{do}(S)}$ are observably equivalent w.r.t $O \setminus S$. Hence, by transitivity, $M_{\text{do}(S)}$ and $\tilde{M}_{\text{do}(S)}$ are observably equivalent w.r.t $O \setminus S$.

□

The reverse implications do not hold in general. Together with Proposition 6.7.6, Proposition 8.2.4 expresses that causal modeling is more refined than probabilistic modeling, and counterfactual modeling is more refined than interventional modeling. This formalizes what Pearl refers to as the "causal hierarchy" or "ladder of causation".

In general, interventional equivalence does not imply counterfactual equivalence. Even interventionally equivalent iSCMs with the same causal mechanism (that differ only in terms of their exogenous distributions) may not be counterfactually equivalent. For example, the iSCMs M^ρ and $M^{\rho'}$ with $\rho \neq \rho'$ in the following example are interventionally equivalent, but not counterfactually equivalent.

Example 8.2.5 (Interventional equivalence does not imply counterfactual equivalence [Daw02]). *For parameter $\rho \in [-1, 1]$, consider the iSCM M^ρ with binary exogenous* *input variable $X \in \{0,1\}$, endogenous variable $Y \in \mathbb{R}$, a single latent exogenous random variable $W = (W_1, W_2) \in \mathbb{R}^2$ with exogenous distribution*

$$\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

and structural equation

$$Y = W_1(1-X) + W_2 X.$$

In a medical setting, this iSCM could be used to model whether a patient was treated or not ($x = 1$ vs. $x = 0$) and the corresponding potential outcome ($Y^{\mathrm{do}(x=1)}$ vs. $Y^{\mathrm{do}(x=0)}$).

Suppose that in the actual world we did not assign treatment to a patient ($x = 0$) and the outcome was $Y^{\mathrm{do}(x=0)} = y \in \mathbb{R}$. Consider the counterfactual query "What would the outcome have been, had we assigned treatment to this patient?". We can answer this question by introducing a parallel counterfactual world in which the exogenous random variables for each patient have the same values as in the actual world, but treatment and outcome may differ. For this, consider the twin iSCM $(M^\rho)^{\mathrm{twin}}$. The counterfactual query then asks for

$$P_{(M^\rho)^{\mathrm{twin}}}((Y')^{\mathrm{do}(x'=1)} \mid Y^{\mathrm{do}(x=0)} = y),$$

where $Y^{\mathrm{do}(x=0)}$ is the factual outcome, and $(Y')^{\mathrm{do}(x'=1)}$ is the counterfactual outcome (which are both marginal potential outcomes of the twinned iSCM). One can calculate that

$$P_{(M^\rho)^{\mathrm{twin}}}\left((Y')^{\mathrm{do}(x'=1)}, Y^{\mathrm{do}(x=0)}\right) = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

and hence $P_{(M^\rho)^{\mathrm{twin}}}((Y')^{\mathrm{do}(x'=1)} \mid Y^{\mathrm{do}(x=0)} = y) = \mathcal{N}(\rho y, 1-\rho^2)$ (by the general formula for conditioning a multivariate Gaussian distribution). Note that the answer to the counterfactual query depends on a quantity ρ that we cannot identify from the Markov kernel $P_{M^\rho}(Y \mid \mathrm{do}(X))$, as it is independent of ρ. Therefore, even unlimited data from a randomized controlled trial would not suffice to determine the value of this particular counterfactual query. Indeed, iSCMs M^ρ and $M^{\rho'}$ with $\rho \neq \rho'$ are interventionally equivalent, but not counterfactually equivalent.

The lesson of this example is that if one attempts to learn an iSCM from data (even from randomized controlled trials with arbitrarily large sample size) it can happen that one still cannot identify the values of some counterfactual probabilities. In other words, data-driven estimation of counterfactual probabilities can be an ill-posed problem. Nevertheless, counterfactual are central in court cases (e.g., to determine responsibility, "the physician treated the patient with drug A and the patient died, would the patient still be alive if the physician had abstained from the treatment?"). The above example shows that one can be on very slippery terrain when it comes to answering such questions.

We have seen that interventionally equivalent iSCMs do not need to have the same graphs. Even counterfactually equivalent iSCMs may have different graphs (see Example 9.5.16).

We have seen that interventional equivalence is preserved under various interventions (Proposition 6.7.8). Analogous statements hold even for counterfactual equivalence.

Proposition 8.2.6. *Assume that M and $\tilde{M}$ are counterfactually equivalent iSCMs w.r.t. $O \subseteq (V \cap \tilde{V})$. Then, for $T \subseteq O \cup (J \cap \tilde{J})$:*

1. *$M_{\mathrm{do}(T)}$ and $\tilde{M}_{\mathrm{do}(T)}$ are counterfactually equivalent w.r.t. $O \setminus T$ (hard interventions with unspecified target value);*
2. *$M_{\mathrm{do}(X_T=\xi_T)}$ and $\tilde{M}_{\mathrm{do}(X_T=\xi_T)}$ are counterfactually equivalent w.r.t. $O \cup T$, for any $\xi_T \in \mathcal{X}_T$ (hard intervention with specified target value);*
3. *$M_{\mathrm{do}(I_T)}$ and $\tilde{M}_{\mathrm{do}(I_T)}$ are counterfactually equivalent w.r.t. O (adding intervention variables).*

Proof. Let $T \subseteq O \cup (J \cap \tilde{J})$. The claims follow straightforwardly from Proposition 8.1.7 and Proposition 6.7.8.

1. To show: $(M_{\mathrm{do}(T)})^{\mathrm{twin}}$ and $(\tilde{M}_{\mathrm{do}(T)})^{\mathrm{twin}}$ are interventionally equivalent w.r.t. $(O \setminus T) \cup (O \setminus T)'$. By Proposition 8.1.7, $(M_{\mathrm{do}(T)})^{\mathrm{twin}} = (M^{\mathrm{twin}})_{\mathrm{do}(T \cup T')}$ and $(\tilde{M}_{\mathrm{do}(T)})^{\mathrm{twin}} = (\tilde{M}^{\mathrm{twin}})_{\mathrm{do}(T \cup T')}$. Since M^{twin} and $\tilde{M}^{\mathrm{twin}}$ are interventionally equivalent iSCMs w.r.t. $O \cup O'$ by assumption, by Proposition 6.7.8, $(M^{\mathrm{twin}})_{\mathrm{do}(T \cup T')}$ and $(\tilde{M}^{\mathrm{twin}})_{\mathrm{do}(T \cup T')}$ are interventionally equivalent w.r.t. $(O \cup O') \setminus (T \cup T')$.

2. Let $\xi_T \in \mathcal{X}_T$. To show: $(M_{\mathrm{do}(X_T=\xi_T)})^{\mathrm{twin}}$ and $(\tilde{M}_{\mathrm{do}(X_T=\xi_T)})^{\mathrm{twin}}$ are interventionally equivalent w.r.t. $(O \cup T) \cup (O \cup T)'$. By Proposition 8.1.7,
$$(M^{\mathrm{twin}})_{\mathrm{do}(X_T=\xi_T, X_{T'}=\xi_T)} = (M_{\mathrm{do}(X_T=\xi_T)})^{\mathrm{twin}}$$
and
$$(\tilde{M}^{\mathrm{twin}})_{\mathrm{do}(X_T=\xi_T, X_{T'}=\xi_T)} = (\tilde{M}_{\mathrm{do}(X_T=\xi_T)})^{\mathrm{twin}}.$$
Since M^{twin} and $\tilde{M}^{\mathrm{twin}}$ are interventionally equivalent iSCMs w.r.t. $O \cup O'$ by assumption, by Proposition 6.7.8, $(M^{\mathrm{twin}})_{\mathrm{do}(X_T=\xi_T, X_{T'}=\xi_T)}$ and $(\tilde{M}^{\mathrm{twin}})_{\mathrm{do}(X_T=\xi_T, X_{T'}=\xi_T)}$ are interventionally equivalent w.r.t. $(O \cup O') \cup (T \cup T')$.

3. To show: $(M_{\mathrm{do}(I_T)})^{\mathrm{twin}}$ and $(\tilde{M}_{\mathrm{do}(I_T)})^{\mathrm{twin}}$ are interventionally equivalent w.r.t. $O \cup O'$. By Proposition 8.1.7,
$$(M^{\mathrm{twin}})_{\mathrm{do}(I_T, I_{T'})} = (M_{\mathrm{do}(I_T)})^{\mathrm{twin}}$$
and
$$(\tilde{M}^{\mathrm{twin}})_{\mathrm{do}(I_T, I_{T'})} = (\tilde{M}_{\mathrm{do}(I_T)})^{\mathrm{twin}}$$
(where we identify I'_t with $I_{t'}$ for all $t \in T$). Since M^{twin} and $\tilde{M}^{\mathrm{twin}}$ are interventionally equivalent iSCMs w.r.t. $O \cup O'$ by assumption, by Proposition 6.7.8, $(M^{\mathrm{twin}})_{\mathrm{do}(I_{T \cup T'})}$ and $(\tilde{M}^{\mathrm{twin}})_{\mathrm{do}(I_{T \cup T'})}$ are interventionally equivalent w.r.t. $O \cup O'$.

□

8.3. Exogenous reparameterizations

Since exogenous random variables are considered latent, certain reparameterizations of those may preserve part of the causal semantics of the observed variables. We will define two types of such exogenous reparameterizations.

8.3.1. Exogenous pushforwards

Definition 8.3.1. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. Let $\tilde{W}$ be a finite index set disjoint from $J \cup V \cup W$, and $\mathcal{X}_{\tilde{W}} = \prod_{\tilde{w} \in \tilde{W}} \mathcal{X}_{\tilde{w}}$ the product of standard measurable spaces $\mathcal{X}_{\tilde{w}}$. Let $\tilde{f} : \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_V$ and $\tilde{\Phi} : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_{\tilde{W}}$ be measurable mappings such that:*

i)

$$f(x_J, x_V, x_W) = \tilde{f}(x_J, x_V, \tilde{\Phi}(x_J, x_W)) \quad M\text{-a.s.},$$

and

ii) there exists a distribution $\tilde{P} = \bigotimes_{\tilde{w} \in \tilde{W}} \tilde{P}_{\tilde{w}}$ with $\tilde{P}_{\tilde{w}} \in \mathcal{P}(\mathcal{X}_{\tilde{w}})$ for $\tilde{w} \in \tilde{W}$ such that:

$$\forall x_J \in \mathcal{X}_J : \quad (\tilde{\Phi}(x_J, \cdot))_* P = \tilde{P}.$$

Then we call the iSCM

$$M_{\mathrm{repar}(\tilde{f}, \tilde{\Phi})} = (J, V, \tilde{W}, \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_{\tilde{W}}, \tilde{P}, \tilde{f})$$

an ***exogenous pushforward*** *of M along $\tilde{\Phi}$.*

Note that given $\tilde{\Phi}$, the distribution $\tilde{P}$ occurring in this definition is unique, but the function $\tilde{f}$ is not necessarily uniquely determined by the definition. Still, the exogenous pushforward $M_{\mathrm{repar}(\tilde{f}, \tilde{\Phi})}$ is well-defined up to equivalence given both $\tilde{\Phi}$ and $\tilde{f}$.

There is some nullset trickery that we need to take care of before we can show various nice properties of this operation on iSCMs.

Lemma 8.3.2. *Let $\tilde{\phi} : \mathcal{X}_W \to \mathcal{X}_{\tilde{W}}$ be a measurable map between two standard measurable spaces. Let $P_{\mathcal{X}_W}$ be a probability measure on $\mathcal{X}_W$ and let $P_{\mathcal{X}_{\tilde{W}}} = P_{\mathcal{X}_W} \circ \tilde{\phi}^{-1}$ be its pushforward under $\tilde{\phi}$. Let $\tilde{\pi} : \mathcal{X}_{\tilde{W}} \to \{0, 1\}$ be a property, i.e., a (measurable) boolean-valued function on $\mathcal{X}_{\tilde{W}}$. Then the property $\pi = \tilde{\pi} \circ \tilde{\phi}$ on $\mathcal{X}_W$ holds $P_{\mathcal{X}_W}$-a.e. if and only if the property $\tilde{\pi}$ holds $P_{\mathcal{X}_{\tilde{W}}}$-a.e..*

Proof. Assume the property $\pi = \tilde{\pi} \circ \tilde{\phi}$ holds $P_{\mathcal{X}_W}$-a.e., then $C = \{x_W \in \mathcal{X}_W : \pi(x_W) = 1\}$ contains a Borel set C^* with $P_{\mathcal{X}_W}$-measure 1, i.e., $C^* \subseteq C$ and $P_{\mathcal{X}_W}(C^*) = 1$. By [Coh13, Proposition 8.2.6], $\tilde{\phi}(C^*)$ is analytic. By [Coh13, Theorem 8.4.1], there exist Borel measurable sets A, B such that $A \subseteq \tilde{\phi}(C^*) \subseteq B$ and $P_{\mathcal{X}_{\tilde{W}}}(A) = P_{\mathcal{X}_{\tilde{W}}}(B)$. Because $\tilde{\phi}$ is measurable, $\tilde{\phi}^{-1}(A)$ and $\tilde{\phi}^{-1}(B)$ are both measurable. Also, $\tilde{\phi}^{-1}(A) \subseteq \tilde{\phi}^{-1}(\tilde{\phi}(C^*)) \subseteq \tilde{\phi}^{-1}(B)$. As $C^* \subseteq \tilde{\phi}^{-1}(\tilde{\phi}(C^*))$, we must have that $P_{\mathcal{X}_W}(\tilde{\phi}^{-1}(B)) \geq P_{\mathcal{X}_W}(C^*) = 1$. Hence

$P_{\mathcal{X}_{\tilde{W}}}(A) = P_{\mathcal{X}_{\tilde{W}}}(B) = 1$. Note that as $C^* \subseteq C$, $A \subseteq \tilde{\phi}(C^*) \subseteq \tilde{\phi}(C) \subseteq \{x_{\tilde{W}} \in \mathcal{X}_{\tilde{W}} : \tilde{\pi}(x_{\tilde{W}}) = 1\}$. Hence the set $\tilde{C} := \{x_{\tilde{W}} \in \mathcal{X}_{\tilde{W}} : \tilde{\pi}(x_{\tilde{W}}) = 1\}$ contains a Borel set of $P_{\mathcal{X}_{\tilde{W}}}$-measure 1, in other words, $\tilde{\pi}$ holds $P_{\mathcal{X}_{\tilde{W}}}$-a.s..

The converse is easier to prove. Suppose $\tilde{C} = \{x_{\tilde{W}} \in \mathcal{X}_{\tilde{W}} : \tilde{\pi}(x_{\tilde{W}}) = 1\}$ contains a Borel set $\tilde{C}^*$ with $P_{\mathcal{X}_{\tilde{W}}}$-measure 1, i.e., $\tilde{C}^* \subseteq \tilde{C}$ and $P_{\mathcal{X}_{\tilde{W}}}(\tilde{C}^*) = 1$. Because $\tilde{\phi}$ is measurable, the set $\tilde{\phi}^{-1}(\tilde{C}^*) \subseteq \tilde{\phi}^{-1}(\tilde{C}) = C$ is measurable and $P_{\mathcal{X}_W}(\tilde{\phi}^{-1}(\tilde{C}^*)) = 1$. □

Corollary 8.3.3. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and $\tilde{M} := M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ an exogenous pushforward of M. Let $\tilde{\pi} : \mathcal{X}_J \times \mathcal{X}_{\tilde{W}} \to \{0, 1\}$ be a binary property and define the property $\pi : \mathcal{X}_J \times \mathcal{X}_W \to \{0, 1\}$ by $\pi(x_J, x_W) := \tilde{\pi}(x_J, \tilde{\Phi}(x_J, x_W))$. Then:*

$$\tilde{\pi}\ \tilde{M}\text{-a.s.} \quad \iff \quad \pi\ M\text{-a.s.}.$$

Proof. By assumption, for all $x_J \in \mathcal{X}_J$, $(\tilde{\Phi}_{x_J})_* P = \tilde{P}$, where the section $\tilde{\Phi}_{x_J} : \mathcal{X}_W \to \mathcal{X}_{\tilde{W}} : x_W \mapsto \tilde{\Phi}(x_J, x_W)$ is measurable.

We want to show that $\tilde{\pi}(x_J, x_{\tilde{W}})$ holds $\tilde{M}$-a.s. if and only if $\tilde{\pi}(x_J, \tilde{\Phi}(x_J, x_W))$ holds M-a.s.. Let $x_J \in \mathcal{X}_J$. Then, the property $\tilde{\pi}_{x_J} := (x_{\tilde{W}} \mapsto \tilde{\pi}(x_J, x_{\tilde{W}}))$ holds $\tilde{P}(X_{\tilde{W}})$-a.s. if and only if $\pi_{x_J} := (x_W \mapsto \tilde{\pi}(x_J, \tilde{\Phi}(x_J, x_W)))$ holds $P(X_W)$-a.s. by Lemma 8.3.2. Since this holds for all $x_J \in \mathcal{X}_J$, this proves the claim. □

Now we can prove our earlier claim that exogenous pushforwards are well-defined up to equivalence.

Lemma 8.3.4. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ an exogenous pushforward of M. Let $\hat{M}$ be an iSCM such that $M \equiv \hat{M}$, and let $\tilde{\hat{f}} : \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_V$ be a measurable mapping such that:*

$$\hat{f}(x_J, x_V, x_W) = \tilde{\hat{f}}(x_J, x_V, \tilde{\Phi}(x_J, x_W)) \quad \hat{M}\text{-a.s.}.$$

Then $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})} \equiv \hat{M}_{\mathrm{repar}(\tilde{\hat{f}},\tilde{\Phi})}$.

Proof. We have $\hat{M} = \left(J, V, W, \mathcal{X}, P, \hat{f}\right)$ and $\hat{M}_{\mathrm{repar}(\tilde{\hat{f}},\tilde{\Phi})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{\hat{f}})$. By assumption, $x_v = f_v(x) \iff x_v = \hat{f}_v(x)$ M-a.s. for all $v \in V$, and furthermore

$$f(x_J, x_V, x_W) = \tilde{f}(x_J, x_V, \tilde{\Phi}(x_J, x_W)) \quad M\text{-a.s.}$$

as well as

$$\hat{f}(x_J, x_V, x_W) = \tilde{\hat{f}}(x_J, x_V, \tilde{\Phi}(x_J, x_W)) \quad M\text{-a.s.},$$

noting that M-a.s. and $\hat{M}$-a.s. mean the same. Hence for $v \in V$,

$$x_v = \tilde{f}_v(x_J, x_V, \tilde{\Phi}(x_J, x_W)) \iff x_v = \tilde{\hat{f}}_v(x_J, x_V, \tilde{\Phi}(x_J, x_W))\ M\text{-a.s.}.$$

By Corollary 8.3.3, this implies

$$x_v = \tilde{f}_v(x) \iff x_v = \tilde{\hat{f}}_v(x)\ M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}\text{-a.s.}.$$

□

Exogenous pushforwards are compatible with hard interventions.

Lemma 8.3.5. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ an exogenous pushforward of M. For an intervention target $T \subseteq V \cup J$:*

$$(M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})})_{\mathrm{do}(T)} = (M_{\mathrm{do}(T)})_{\mathrm{repar}(\tilde{f}_{\setminus T},\tilde{\Phi})}.$$

Proof. This follows by writing out the definitions. □

The twinning operation is only compatible with exogenous pushforwards under restrictive assumptions.

Lemma 8.3.6. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ an exogenous pushforward of M. If $\tilde{\Phi}$ is constant in X_J, that is, $\tilde{\Phi}(x_J, x_W) = \tilde{\Phi}(\tilde{x}_J, x_W)$ for all $x_W \in \mathcal{X}_W$ and all $x_J, \tilde{x}_J \in \mathcal{X}_J$, then:*

$$(M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})})^{\mathrm{twin}} = (M^{\mathrm{twin}})_{\mathrm{repar}(\tilde{f}^{\mathrm{twin}},\tilde{\Phi})}.$$

Proof. This follows by writing out the definitions. In particular, we check that

$$\begin{aligned} f^{\mathrm{twin}}(x_J, x_{J'}, x_V, x_{V'}, x_W) &= (f(x_J, x_V, x_W), f(x_{J'}, x_{V'}, x_W)) \\ &= (\tilde{f}(x_J, x_V, \tilde{\Phi}(x_W)), \tilde{f}(x_{J'}, x_{V'}, \tilde{\Phi}(x_W)) \\ &= \tilde{f}^{\mathrm{twin}}(x_J, x_{J'}, x_V, x_{V'}, \tilde{\Phi}(x_W)). \end{aligned}$$

For the last equality to hold, it suffices that $\tilde{\Phi}$ does not depend on X_J. □

We can pull back partial solution functions from exogenous pushforwards.

Lemma 8.3.7. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ an exogenous pushforward of M. If $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$ has partial solution function $\tilde{g}^{[L]} : \mathcal{X}_J \times \mathcal{X}_{V \setminus L} \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_L$ w.r.t. $L \subseteq V$, then M has partial solution function*

$$g^{[L]} : \mathcal{X}_J \times \mathcal{X}_{V \setminus L} \times \mathcal{X}_W \to \mathcal{X}_L : (x_J, x_{V \setminus L}, x_W) \mapsto \tilde{g}^{[L]}(x_J, x_{V \setminus L}, \tilde{\Phi}(x_J, x_W))$$

w.r.t. L. If $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$ is essentially uniquely solvable w.r.t. $L \subseteq V$, then so is M.

Proof. If $\tilde{M} := M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$ has partial solution function $\tilde{g}^{[L]}$ w.r.t. $L \subseteq V$, then

$$x_L = \tilde{f}_L(x_J, x_V, x_{\tilde{W}}) \impliedby x_L = \tilde{g}^{[L]}(x_J, x_{V \setminus L}, x_{\tilde{W}}) \quad \tilde{M}\text{-a.s.}.$$

By Corollary 8.3.3,

$$x_L = \tilde{f}_L(x_J, x_V, \tilde{\Phi}(x_J, x_W)) \impliedby x_L = \tilde{g}^{[L]}(x_J, x_{V \setminus L}, \tilde{\Phi}(x_J, x_W)) \quad M\text{-a.s.}.$$

By assumption,

$$f(x_J, x_V, x_W) = \tilde{f}(x_J, x_V, \tilde{\Phi}(x_J, x_W)) \quad M\text{-a.s.}.$$

Hence

$$x_L = f_L(x_J, x_V, x_W) \impliedby x_L = \tilde{g}^{[L]}(x_J, x_{V \setminus L}, \tilde{\Phi}(x_J, x_W)) \quad M\text{-a.s.}.$$

Therefore, M has partial solution function $g^{[L]}$ w.r.t. L. If $\tilde{M}$ is essentially uniquely solvable w.r.t. L, then all logical implications above become logical equivalences, and hence M is essentially uniquely solvable w.r.t. L as well. □

Theorem 8.3.8. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ an exogenous pushforward of M. If both M and $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$ are simple, then M is observably and interventionally equivalent to $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$ w.r.t. V. If, furthermore, $\tilde{\Phi}$ does not depend on X_J, then M is even counterfactually equivalent to $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$ w.r.t. V.*

Proof. Let $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ be a (essentially unique) solution function of M, and $\tilde{g} : \mathcal{X}_J \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_V$ a (essentially unique) solution function of $\tilde{M} := M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$. By Lemma 8.3.7 and the essential uniqueness of the solution functions, we get

$$g = \tilde{g} \circ (\mathrm{id}_{\mathcal{X}_J}, \tilde{\Phi}).$$

We now show that the marginal Markov kernels $P_M(X_V \mid \mathrm{do}(X_J))$ and $P_{M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}}(X_V \mid \mathrm{do}(X_J))$ are the same. By Theorem 6.6.4,

$$\begin{aligned}
P_M(X_V \mid \mathrm{do}(X_J)) &= (g)_* \left(\delta(X_J|X_J) \otimes P(X_W)\right) \\
&= (\tilde{g} \circ (\mathrm{id}_{\mathcal{X}_J}, \tilde{\Phi}))_* \left(\delta(X_J|X_J) \otimes P(X_W)\right) \\
&= (\tilde{g})_* \left(\delta(X_J|X_J) \otimes \tilde{P}(X_{\tilde{W}})\right) \\
&= P_{M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}}(X_V \mid \mathrm{do}(X_J)).
\end{aligned}$$

This shows the observable equivalence w.r.t. V.

Let $T \subseteq V$. Then $(M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})})_{\mathrm{do}(T)} = (M_{\mathrm{do}(T)})_{\mathrm{repar}(\tilde{f}_{\setminus T},\tilde{\Phi})}$ by Lemma 8.3.5. The observable equivalence of $M_{\mathrm{do}(T)}$ and $(M_{\mathrm{do}(T)})_{\mathrm{repar}(\tilde{f}_{\setminus T},\tilde{\Phi})}$ w.r.t. $V \setminus T$ hence implies the observable equivalence of $M_{\mathrm{do}(T)}$ and $(M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})})_{\mathrm{do}(T)}$ w.r.t. $V \setminus T$. Since this holds for all $T \subseteq V$, M and $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$ are interventionally equivalent w.r.t. V.

Assume now that $\tilde{\Phi}$ does not depend on X_J. By Lemma 8.3.6, $(M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})})^{\mathrm{twin}} = (M^{\mathrm{twin}})_{\mathrm{repar}(\tilde{f}^{\mathrm{twin}},\tilde{\Phi})}$. Since M^{twin} and its exogenous pushforward $(M^{\mathrm{twin}})_{\mathrm{repar}(\tilde{f}^{\mathrm{twin}},\tilde{\Phi})}$ are interventionally equivalent w.r.t. $V \cup V'$, M and $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$ are counterfactually equivalent w.r.t. V. □

Proposition 8.3.9. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ an exogenous pushforward of M, where $\mathcal{X}_J$ and $\mathcal{X}_{\tilde{W}}$ are discrete spaces. Then M is simple if and only if $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$ is simple.*

Proof. Denote $\tilde{M} := M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$. The implication that M is simple if $\tilde{M}$ is simple follows from Lemma 8.3.7, so we only have to prove the converse. Assume that M is simple. To show that $\tilde{M}$ is simple, by Remark 6.6.6 it suffices to show that for each $L \subseteq V$

$$\exists! x_L \in \mathcal{X}_L : x_L = \tilde{f}_L(x_J, x_{V \setminus L}, x_L, x_{\tilde{W}})$$

up to a measurable $\tilde{M}$-null set. Because any $\tilde{M}$-null set is measurable because of the discreteness of $\mathcal{X}_J$ and $\mathcal{X}_{\tilde{W}}$, it suffices if this holds up to a $\tilde{M}$-null set.

Let $L \subseteq V$ and consider the property which expresses that the structural equations of $\tilde{M}^{[L]}$ have a unique solution:

$$\tilde{\pi}(x_J, x_{\tilde{W}}) := \left[\forall x_{V \setminus L} \in \mathcal{X}_{V \setminus L} \exists! x_L \in \mathcal{X}_L : x_L = \tilde{f}_L(x_J, x_{V \setminus L}, x_L, x_{\tilde{W}})\right].$$

Then the corresponding property for M is:

$$\pi(x_J, x_W) = \left[\forall x_{V\setminus L} \in \mathcal{X}_{V\setminus L} \exists! x_L \in \mathcal{X}_L : x_L = \tilde{f}_L(x_J, x_{V\setminus L}, x_L, \tilde{\Phi}(x_J, x_W))\right],$$

for which we have that $\tilde{\pi}$ $\tilde{M}$-a.s. $\iff$ π M-a.s. by Corollary 8.3.3. Since

$$f(x_J, x_V, x_W) = \tilde{f}(x_J, x_V, \tilde{\Phi}(x_J, x_W)) \quad M\text{-a.s.}$$

by assumption, we conclude:

$$\begin{aligned}
&\forall x_{V\setminus L} \in \mathcal{X}_{V\setminus L} \exists! x_L \in \mathcal{X}_L : x_L = \tilde{f}_L(x_J, x_V, x_{\tilde{W}}) \quad \tilde{M}\text{-a.s.} \\
\iff &\forall x_{V\setminus L} \in \mathcal{X}_{V\setminus L} \exists! x_L \in \mathcal{X}_L : x_L = \tilde{f}_L(x_J, x_V, \tilde{\Phi}(x_J, x_W)) \quad M\text{-a.s.} \\
\iff &\forall x_{V\setminus L} \in \mathcal{X}_{V\setminus L} \exists! x_L \in \mathcal{X}_L : x_L = f_L(x_J, x_V, x_W) \quad M\text{-a.s..}
\end{aligned}$$

Since M is simple, it is essentially uniquely solvable w.r.t. L, and hence the last condition holds. Therefore, also the first condition holds. This proves the claim. □

An exogenous pushforward can be inverted if the reparameterization mapping $\tilde{\Phi}$ has an 'essential left-inverse'.

Lemma 8.3.10. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and $\tilde{M} := M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ an exogenous pushforward of M. Suppose there exists an essential left-inverse of $\tilde{\Phi}$, that is, a measurable function $\Psi : \mathcal{X}_J \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_W$ such that $\Psi(x_J, \tilde{\Phi}(x_J, x_W)) = x_W$ M-a.s.. Then $M \equiv \tilde{M}_{\mathrm{repar}(f,\Psi)}$.*

Proof. Since Ψ is an essential left-inverse of $\tilde{\Phi}$,

$$(\mathrm{id}_{\mathcal{X}_J}, \Psi) \circ (\mathrm{id}_{\mathcal{X}_J}, \tilde{\Phi}) = (\mathrm{id}_{\mathcal{X}_J}, \mathrm{id}_{\mathcal{X}_W}) \ M\text{-a.s..}$$

By definition,

$$\tilde{M}_{\mathrm{repar}(\hat{f},\Psi)} = (J, V, W, \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_W, \hat{P}, \hat{f})$$

with $\hat{f}$ such that

$$\tilde{f}(x_J, x_V, x_{\tilde{W}}) = \hat{f}(x_J, x_V, \Psi(x_J, x_{\tilde{W}})) \quad \tilde{M}\text{-a.s.}$$

and for all $x_J \in \mathcal{X}_J$,

$$(\Psi(x_J, \cdot))_* \tilde{P} = \hat{P}$$

We show that picking $\hat{f} = f$ works, and that $\hat{P} = P$.

If $\hat{f} = f$, then

$$\tilde{f}(x_J, x_V, \tilde{\Phi}(x_J, x_W)) = \hat{f}(x_J, x_V, x_W) = \hat{f}(x_J, x_V, \Psi(x_J, \tilde{\Phi}(x_J, x_W))) \quad M\text{-a.s.}$$

Hence, by Corollary 8.3.3,

$$\tilde{f}(x_J, x_V, x_{\tilde{W}}) = \hat{f}(x_J, x_V, \Psi(x_J, x_{\tilde{W}})) \quad \tilde{M}\text{-a.s.}$$

which is precisely what was required.

For all $x_J \in \mathcal{X}_J$,

$$\begin{aligned}
\hat{P} &= (\Psi(x_J,\cdot))_*\tilde{P} \\
&= (\Psi(x_J,\cdot))_*\big(\tilde{\Phi}(x_J,\cdot))_*P\big) \\
&= \big(P \circ \tilde{\Phi}(x_J,\cdot)^{-1}\big) \circ \Psi(x_J,\cdot)^{-1} \\
&= P \circ \big(\tilde{\Phi}(x_J,\cdot)^{-1} \circ \Psi(x_J,\cdot)^{-1}\big) \\
&= P \circ \big(\Psi(x_J,\cdot) \circ \tilde{\Phi}(x_J,\cdot)\big)^{-1} \\
&= P
\end{aligned}$$

So that also works out. □

One example of an exogenous pushforward of this type that one encounters occassionally is the operation of merging exogenous random variables.

Example 8.3.11 (Merging exogenous random variables)**.** *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. Let $\tilde{W}$ be a partition of W. Take for $\mathcal{X}_{\tilde{W}} = \prod_{\tilde{w}\in\tilde{W}} \mathcal{X}_{\tilde{w}}$ the product of standard measurable spaces $\mathcal{X}_{\tilde{w}} = \prod_{w\in\tilde{w}} \mathcal{X}_w$. Let $\tilde{\Phi} : \mathcal{X}_W \to \mathcal{X}_{\tilde{W}}$ be the natural identification that maps $x_W = (x_w)_{w\in W} \in \mathcal{X}_W$ with components $x_w \in \mathcal{X}_w$ to $x_{\tilde{W}} = (x_{\tilde{w}})_{\tilde{w}\in\tilde{W}} \in \mathcal{X}_{\tilde{W}}$ with components $x_{\tilde{w}} = (x_w)_{w\in\tilde{w}} \in \mathcal{X}_{\tilde{w}}$. Because $\tilde{W}$ is a partition of W, $\tilde{\Phi}$ is a bijection. Take*

$$\tilde{f} : \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_V : (x_J, x_V, x_{\tilde{W}}) \mapsto f(x_J, x_V, \tilde{\Phi}^{-1}(x_{\tilde{W}})).$$

It is immediate from Definition 8.3.1 that

$$M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})} = (J, V, \tilde{W}, \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_{\tilde{W}}, \tilde{P}, \tilde{f})$$

is an exogenous pushforward of M. Because $\tilde{\Phi}$ is independent of x_J and bijective, $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$ is observably, interventionally and counterfactually equivalent to M w.r.t. V in case M is simple.

The following example shows that an exogenous pushforward need not be counterfactually equivalent w.r.t. V if $\tilde{\Phi}$ depends on X_J.

Example 8.3.12. *Consider the acyclic iSCM M with exogenous input variable X_1 with co-domain $\{-1, 1\}$, endogenous variables X_2, X_3 with co-domains $\{-1, 1\}$, $\{-2, 0, 2\}$, respectively, and causal mechanism*

$$\begin{aligned}
X_2 &= f_2(X_1, X_B) = X_1 X_B \\
X_3 &= f_3(X_2, X_B) = X_2 + X_B,
\end{aligned}$$

with exogenous random variable $X_B \sim \mathrm{Uni}(\{-1, 1\})$. Consider the exogenous pushforward $\tilde{M}$ with exogenous random variable $X_{\tilde{B}}$ with co-domain $\{-1, 1\}$, mapping $\tilde{\Phi} : \{-1, 1\}^2 \to \{-1, 1\} : (x_1, x_B) \mapsto x_1 x_B$, and causal mechanism

$$\begin{aligned}
X_2 &= \tilde{f}_2(X_{\tilde{B}}) = X_{\tilde{B}} \\
X_3 &= \tilde{f}_3(X_1, X_2, X_{\tilde{B}}) = X_2 + X_1 X_{\tilde{B}}.
\end{aligned}$$

Its exogenous distribution is $\tilde{\Phi}_*(\mathbb{P}(X_B)) = \mathrm{Uni}(\{-1,1\}) = \mathbb{P}(X_{\tilde{B}})$. *It is indeed an exogenous pushforward:*

$$\begin{aligned}\tilde{f}(x_1, x_2, \tilde{\Phi}(x_1, x_B)) &= (\tilde{\Phi}(x_1, x_B), x_2 + x_1\tilde{\Phi}(x_1, x_B)) \\ &= (x_1 x_B, x_2 + x_1 x_1 x_B) \\ &= (x_1 x_B, x_2 + x_B) \\ &= f(x_1, x_2, x_B).\end{aligned}$$

Both M and $\tilde{M}$ are acyclic, hence simple. By Theorem 8.3.8, $\tilde{M}$ is observably and interventionally equivalent to M w.r.t. $\{X_2, X_3\}$. However, $\tilde{M}$ is not counterfactually equivalent to M w.r.t. $\{X_2, X_3\}$, as one can check explicitly.

8.3.2. Exogenous pullback

Since exogenous random variables are considered latent, certain reparameterizations of those may preserve part of the causal semantics of the observed variables.

Definition 8.3.13. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM. Let $\tilde{W}$ be a finite index set disjoint from $J \cup V \cup W$, and $\mathcal{X}_{\tilde{W}} = \prod_{\tilde{w}\in\tilde{W}} \mathcal{X}_{\tilde{w}}$ the product of standard measurable spaces $\mathcal{X}_{\tilde{w}}$. Let $\Phi : \mathcal{X}_J \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_W$ be a measurable mapping. Let $\tilde{P} = \bigotimes_{\tilde{w}\in\tilde{W}} \tilde{P}_{\tilde{w}}$ with $\tilde{P}_{\tilde{w}} \in \mathcal{P}(\mathcal{X}_{\tilde{w}})$ for all $\tilde{w} \in \tilde{W}$. Define the function $\tilde{f} : \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_V$ by:*

$$\tilde{f}(x_J, x_V, x_{\tilde{W}}) = f(x_J, x_V, \Phi(x_J, x_{\tilde{W}}))$$

for all $x_J \in \mathcal{X}_J, x_V \in \mathcal{X}_V, x_{\tilde{W}} \in \mathcal{X}_{\tilde{W}}$. If

$$\forall x_J \in \mathcal{X}_J : \quad (\Phi(x_J, \cdot))_* \tilde{P} = P,$$

we call the iSCM

$$M_{\mathrm{repar}(\Phi,\tilde{P})} = (J, V, \tilde{W}, \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_{\tilde{W}}, \tilde{P}, \tilde{f})$$

an ***exogenous pullback*** *of M along Φ.*

Note that given Φ, the function $\tilde{f}$ occurring in this definition is unique, but the distribution $\tilde{f}$ is not necessarily uniquely determined by the definition. Still, the exogenous pullback $M_{\mathrm{repar}(\Phi,\tilde{P})}$ is well-defined given both Φ and $\tilde{P}$.

There is again some nullset trickery that we need to take care of before we can show various nice properties of this operation on iSCMs.

Lemma 8.3.14. *Let $\phi : \mathcal{X}_{\tilde{W}} \to \mathcal{X}_W$ be a measurable map between two standard measurable spaces. Let $P_{\mathcal{X}_{\tilde{W}}}$ be a probability measure on $\mathcal{X}_{\tilde{W}}$ and let $P_{\mathcal{X}_W} = P_{\mathcal{X}_{\tilde{W}}} \circ \phi^{-1}$ be its pushforward under ϕ. Let $\pi : \mathcal{X}_W \to \{0,1\}$ be a property, i.e., a (measurable) boolean-valued function on $\mathcal{X}_W$. Then the property $\tilde{\pi} = \pi \circ \phi$ on $\mathcal{X}_{\tilde{W}}$ holds $P_{\mathcal{X}_{\tilde{W}}}$-a.e. if and only if the property π holds $P_{\mathcal{X}_W}$-a.e..*

Proof. This is just a reformulation of Lemma 8.3.2. ☐

Corollary 8.3.15. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and $\tilde{M} := M_{\text{repar}(\Phi,\tilde{P})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ an exogenous pullback of M. Let $\pi : \mathcal{X}_J \times \mathcal{X}_W \to \{0, 1\}$ be a binary property and define the property $\tilde{\pi} : \mathcal{X}_J \times \mathcal{X}_{\tilde{W}} \to \{0, 1\}$ by $\tilde{\pi}(x_J, x_{\tilde{W}}) := \pi(x_J, \Phi(x_J, x_{\tilde{W}}))$. Then:*

$$\pi \ M\text{-a.s.} \quad \Longleftrightarrow \quad \tilde{\pi} \ \tilde{M}\text{-a.s.}.$$

Proof. By assumption, for all $x_J \in \mathcal{X}_J$, $(\Phi_{x_J})_* \tilde{P} = P$, where the section $\Phi_{x_J} : \mathcal{X}_{\tilde{W}} \to \mathcal{X}_W : x_{\tilde{W}} \mapsto \Phi(x_J, x_{\tilde{W}})$ is measurable.

We want to show that $\pi(x_J, x_W)$ holds M-a.s. if and only if $\pi(x_J, \Phi(x_J, x_{\tilde{W}}))$ holds $\tilde{M}$-a.s.. Let $x_J \in \mathcal{X}_J$. Then, the property $\pi_{x_J} := (x_W \mapsto \pi(x_J, x_W))$ holds $P(X_W)$-a.s. if and only if $\tilde{\pi}_{x_J} := (x_{\tilde{W}} \mapsto \pi(x_J, \Phi_{x_J}(x_{\tilde{W}})))$ holds $P(X_{\tilde{W}})$-a.s. by Lemma 8.3.14. Since this holds for all $x_J \in \mathcal{X}_J$, this proves the claim. □

Lemma 8.3.16. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and $M_{\text{repar}(\Phi,\tilde{P})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ an exogenous pullback of M. Let $\hat{M}$ be an iSCM such that $M \equiv \hat{M}$. Then $M_{\text{repar}(\Phi,\tilde{P})} \equiv \hat{M}_{\text{repar}(\hat{\Phi},\tilde{P})}$.*

Proof. We have $\hat{M} = (J, V, W, \mathcal{X}, P, \hat{f})$ and $\hat{M}_{\text{repar}(\Phi,\tilde{P})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{\hat{f}})$.

By assumption, $x_v = f_v(x) \iff x_v = \hat{f}_v(x)$ M-a.s. for all $v \in V$. By Corollary 8.3.15, this implies that for $v \in V$,

$$x_v = f_v(x_J, x_V, \Phi(x_J, x_{\tilde{W}})) \iff x_v = \hat{f}_v(x_J, x_V, \Phi(x_J, x_{\tilde{W}})) \ M_{\text{repar}(\Phi,\tilde{P})}\text{-a.s.}$$

By the definition of exogenous pullbacks, we can rewrite this as

$$x_v = \tilde{f}_v(x_J, x_V, x_{\tilde{W}}) \iff x_v = \tilde{\hat{f}}_v(x_J, x_V, x_{\tilde{W}}) \ M_{\text{repar}(\Phi,\tilde{P})}\text{-a.s.}$$

□

Exogenous pullbacks are compatible with hard interventions.

Lemma 8.3.17. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and $M_{\text{repar}(\Phi,\tilde{P})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ an exogenous pullback of M. For an intervention target $T \subseteq V \cup J$:*

$$(M_{\text{repar}(\Phi,\tilde{P})})_{\text{do}(T)} = (M_{\text{do}(T)})_{\text{repar}(\Phi,\tilde{P})}.$$

Proof. This follows by writing out the definitions. □

The twinning operation is only compatible with exogenous pullbacks under restrictive assumptions.

Lemma 8.3.18. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and $M_{\text{repar}(\Phi,\tilde{P})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ an exogenous pullback of M. If Φ is constant in X_J, that is, $\Phi(x_J, x_{\tilde{W}}) = \Phi(\tilde{x}_J, x_{\tilde{W}}) =: \Phi(x_{\tilde{W}})$ for all $x_{\tilde{W}} \in \mathcal{X}_{\tilde{W}}$ and all $x_J, \tilde{x}_J \in \mathcal{X}_J$, then:*

$$(M_{\text{repar}(\Phi,\tilde{P})})^{\text{twin}} = (M^{\text{twin}})_{\text{repar}(\Phi,\tilde{P})}.$$

Proof. This follows by writing out the definitions. In particular, we check that

$$\begin{aligned}\tilde{f}^{\text{twin}}(x_J, x_{J'}, x_V, x_{V'}, x_{\tilde{W}}) &= (\tilde{f}(x_J, x_V, x_{\tilde{W}}), \tilde{f}(x_{J'}, x_{V'}, x_{\tilde{W}})) \\ &= (f(x_J, x_V, \Phi(x_J, x_{\tilde{W}})), f(x_{J'}, x_{V'}, \Phi(x_{J'}, x_{\tilde{W}}))) \\ &= (f(x_J, x_V, \Phi(x_{\tilde{W}})), f(x_{J'}, x_{V'}, \Phi(x_{\tilde{W}}))) \\ &= f^{\text{twin}}(x_J, x_{J'}, x_V, x_{V'}, \Phi(x_{\tilde{W}})).\end{aligned}$$

For the last equality to hold, it suffices that Φ does not depend on X_J. □

We can push forward partial solution functions to exogenous pullbacks.

Lemma 8.3.19. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and $M_{\text{repar}(\Phi,\tilde{P})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ an exogenous pullback of M. If M has partial solution function $g^{[L]} : \mathcal{X}_J \times \mathcal{X}_{V\setminus L} \times \mathcal{X}_W \to \mathcal{X}_L$ w.r.t. $L \subseteq V$, then $M_{\text{repar}(\Phi,\tilde{P})}$ has partial solution function*

$$\tilde{g}^{[L]} : \mathcal{X}_J \times \mathcal{X}_{V\setminus L} \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_L : (x_J, x_{V\setminus L}, x_{\tilde{W}}) \mapsto g^{[L]}(x_J, x_{V\setminus L}, \Phi(x_J, x_{\tilde{W}}))$$

w.r.t. L. If M is essentially uniquely solvable w.r.t. $L \subseteq V$, then so is $M_{\text{repar}(\Phi,\tilde{P})}$.

Proof. If M has partial solution function $g^{[L]}$ w.r.t. $L \subseteq V$, then

$$x_L = f_L(x_J, x_V, x_W) \impliedby x_L = g^{[L]}(x_J, x_{V\setminus L}, x_W) \quad M\text{-a.s..}$$

By Corollary 8.3.15,

$$x_L = f_L(x_J, x_V, \Phi(x_J, x_{\tilde{W}})) \impliedby x_L = g^{[L]}(x_J, x_{V\setminus L}, \Phi(x_J, x_{\tilde{W}})) \quad M_{\text{repar}(\Phi,\tilde{P})}\text{-a.s..}$$

Hence

$$x_L = \tilde{f}_L(x_J, x_V, x_{\tilde{W}}) \impliedby x_L = \tilde{g}^{[L]}(x_J, x_{V\setminus L}, x_{\tilde{W}}) \quad M_{\text{repar}(\Phi,\tilde{P})}\text{-a.s..}$$

Therefore, $M_{\text{repar}(\Phi,\tilde{P})}$ has partial solution function $\tilde{g}^{[L]}$ w.r.t. L. If M is essentially uniquely solvable w.r.t. L, then all logical implications above become logical equivalences, and then $M_{\text{repar}(\Phi,\tilde{P})}$ is essentially uniquely solvable w.r.t. L as well. □

Theorem 8.3.20. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be an iSCM, and $M_{\text{repar}(\Phi,\tilde{P})} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ an exogenous pullback of M. If M is simple, then so is $M_{\text{repar}(\Phi,\tilde{P})}$, and M is observably and interventionally equivalent to $M_{\text{repar}(\Phi,\tilde{P})}$ w.r.t. V. If, furthermore, Φ does not depend on X_J, then M is even counterfactually equivalent to $M_{\text{repar}(\Phi,\tilde{P})}$ w.r.t. V.*

Proof. If M is simple, then it is essentially uniquely solvable w.r.t. T for all $T \subseteq V$. By Lemma 8.3.19, $M_{\text{repar}(\Phi,\tilde{P})}$ is then essentially uniquely solvable w.r.t. T for all $T \subseteq V$. Hence $M_{\text{repar}(\Phi,\tilde{P})}$ is simple.

Let $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ be a solution function of M. Define the function $\tilde{g} : \mathcal{X}_J \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_V$ by

$$\tilde{g} = g \circ (\text{id}_{\mathcal{X}_J}, \Phi).$$

By Lemma 8.3.19, $\tilde{g}$ is a solution function of $M_{\text{repar}(\Phi,\tilde{P})}$.

We now show that the marginal Markov kernels $P_M(X_V \mid \text{do}(X_J))$ and $P_{M_{\text{repar}(\Phi,\tilde{P})}}(X_V \mid \text{do}(X_J))$ are the same. By Theorem 6.6.4,

$$\begin{aligned}
P_{M_{\text{repar}(\Phi,\tilde{P})}}(X_V \mid \text{do}(X_J)) &= (\tilde{g})_* \left(\delta(X_J|X_J) \otimes \tilde{P}(X_{\tilde{W}})\right) \\
&= (g \circ (\text{id}_{\mathcal{X}_J}, \Phi))_* \left(\delta(X_J|X_J) \otimes \tilde{P}(X_{\tilde{W}})\right) \\
&= (g)_* \left(\delta(X_J|X_J) \otimes P(X_W)\right) \\
&= P_M(X_V \mid \text{do}(X_J)).
\end{aligned}$$

This shows the observable equivalence w.r.t. V.

Let $T \subseteq V$. Then $(M_{\text{repar}(\Phi,\tilde{P})})_{\text{do}(T)} = (M_{\text{do}(T)})_{\text{repar}(\Phi,\tilde{P})}$ by Lemma 8.3.17. The observable equivalence of $M_{\text{do}(T)}$ and $(M_{\text{do}(T)})_{\text{repar}(\Phi,\tilde{P})}$ w.r.t. $V \setminus T$ hence implies the observable equivalence of $M_{\text{do}(T)}$ and $(M_{\text{repar}(\Phi,\tilde{P})})_{\text{do}(T)}$ w.r.t. $V \setminus T$. Since this holds for all $T \subseteq V$, M and $M_{\text{repar}(\Phi,\tilde{P})}$ are interventionally equivalent w.r.t. V.

Assume now that Φ does not depend on X_J. By Lemma 8.3.18, $(M_{\text{repar}(\Phi,\tilde{P})})^{\text{twin}} = (M^{\text{twin}})_{\text{repar}(\Phi,\tilde{P})}$. Since M^{twin} and its exogenous pullback $(M^{\text{twin}})_{\text{repar}(\Phi,\tilde{P})}$ are interventionally equivalent w.r.t. $V \cup V'$, M and $M_{\text{repar}(\Phi,\tilde{P})}$ are counterfactually equivalent w.r.t. V. □

Example 8.3.11 showed that merging exogenous random variables can be done via an exogenous pushforward. Since the reparameterization mapping between the exogenous random state spaces is invertible, this operation can also be seen as an exogenous pullback.

Example 8.3.21 (Merging exogenous random variables). *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let $\tilde{W}$ be a partition of W. Let $\mathcal{X}_{\tilde{w}} = \prod_{w \in \tilde{w}} \mathcal{X}_w$ for each subset $\tilde{w} \in \tilde{W}$. Let $\Phi : \mathcal{X}_{\tilde{W}} \to \mathcal{X}_W$ be the canonical projections. Take $\tilde{P} = \bigotimes_{\tilde{w} \in \tilde{W}} \tilde{P}_{\tilde{w}}$ with $\tilde{P}_{x_{\tilde{w}}}(X_{\tilde{w}}) = P(X_{\tilde{w}})$ for all $\tilde{w} \in \tilde{W}$, and define $\tilde{f} : \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_V$ by*

$$\tilde{f}(x_J, x_V, x_{\tilde{W}}) = f(x_J, x_V, \Phi(x_{\tilde{W}})).$$

It is immediate from Definition 8.3.13 that

$$M_{\text{repar}(\Phi,\tilde{P})} = (J, V, \tilde{W}, \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_{\tilde{W}}, \tilde{P}, \tilde{f})$$

Then $M_{\text{repar}(\Phi,\tilde{P})}$ is observably, interventionally and counterfactually equivalent to M w.r.t. V.

Example 8.3.22 (Reparameterization trick). *The reparameterization trick: suppose we have an iSCM M with kernel $P(X_W \mid \text{do}(X_J))$ (where now we allow a dependence on X_J!) and a structural equation*

$$X_V = f(X_J, X_V, X_W).$$

Let $\tilde{W} := \{\tilde{w}\}$ for some index $\tilde{w}$ not in $J \cup V \cup W$, $\mathcal{X}_{\tilde{w}} := [0,1]$, $\tilde{P}(X_{\tilde{W}})$ the uniform distribution on $\mathcal{X}_{\tilde{W}}$. Theorem 2.7.3 with Remark 2.7.4 states that there exists a function $\tilde{R} : \mathcal{X}_J \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_W$ such that:

$$P(X_W | \operatorname{do}(X_J)) = \delta(\tilde{R}|X_{\tilde{W}}, X_J) \circ \tilde{P}(X_{\tilde{W}}).$$

Consider now $\tilde{M} := M_{\operatorname{repar}(\tilde{R},\tilde{P})}$, the exogenous pullback of M along $\tilde{R}$. It has causal mechanism $\tilde{f} : \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_V$ defined by

$$\tilde{f}(x_J, x_V, x_{\tilde{W}}) = f(x_J, x_V, \tilde{R}(x_J, x_{\tilde{W}}))$$

for all $x_J \in \mathcal{X}_J, x_V \in \mathcal{X}_V, x_{\tilde{W}} \in \mathcal{X}_{\tilde{W}}$. If $\tilde{R}$ depends on X_J, then $\tilde{M}$ may not be counterfactually equivalent to M w.r.t. V.

Note that we can also consider M to be an exogenous pushforward of $\tilde{M}$: $M = \tilde{M}_{\operatorname{repar}(f,\tilde{R})}$.

8.3.3. Pushforward or pullback?

We have defined two dual types of exogenous reparameterizations: exogenous pushforwards and exogenous pullbacks. The former specifies the target mechanism and forward reparameterization function, whereas the latter specifies the target distribution and the backward reparameterization function. Sometimes, an operation is most naturally specified as an exogenous pushforward (for example, the response variable reparameterization that we discuss in Section 8.4). In other cases, an operation is most naturally specified as an exogenous pullback (for example, the reparameterization trick of Example 8.3.22).

We summarize both definitions:

Definition 8.3.23. *Let M and $\tilde{M}$ be iSCMs. Let $M = (J, V, W, \mathcal{X}, P, f)$ and $\tilde{M} = (J, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ be iSCMs. We call $\tilde{M}$ an **exogenous pushforward of M through measurable function** $\tilde{\Phi} : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_{\tilde{W}}$ if*

1. $f = \tilde{f} \circ (\mathrm{id}_{\mathcal{X}_J}, \mathrm{id}_{\mathcal{X}_V}, \tilde{\Phi}) = (\mathrm{id}_{\mathcal{X}_J}, \mathrm{id}_{\mathcal{X}_V}, \tilde{\Phi})^* \tilde{f}$ *("pullback of $\tilde{f}$ by $(\mathrm{id}_{\mathcal{X}_J}, \mathrm{id}_{\mathcal{X}_V}, \tilde{\Phi})$")*
2. $\tilde{P} = (\mathrm{id}_{\mathcal{X}_J}, \tilde{\Phi})_* P$ *("pushforward of P by $(\mathrm{id}_{\mathcal{X}_J}, \tilde{\Phi})$")*

*We call $\tilde{M}$ an **exogenous pullback of M through measurable function** $\Phi : \mathcal{X}_J \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_W$ if*

1. $\tilde{f} = f \circ (\mathrm{id}_{\mathcal{X}_J}, \mathrm{id}_{\mathcal{X}_V}, \Phi) = (\mathrm{id}_{\mathcal{X}_V}, \Phi)^* f$ *("pullback of f by $(\mathrm{id}_{\mathcal{X}_J}, \mathrm{id}_{\mathcal{X}_V}, \Phi)$")*
2. $P = (\mathrm{id}_{\mathcal{X}_J}, \Phi)_* \tilde{P}$ *("pushforward of $\tilde{P}$ by $(\mathrm{id}_{\mathcal{X}_J}, \Phi)$")*

It follows immediately that

Proposition 8.3.24. *Let $M = (J, V, W, \mathcal{X}, P, f)$ and $\tilde{M} = (\tilde{J}, V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f})$ be iSCMs. Let $\Phi : \mathcal{X}_{\tilde{J}} \times \mathcal{X}_{\tilde{W}} \to \mathcal{X}_J \times \mathcal{X}_W$ and $\tilde{\Phi} : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_{\tilde{J}} \times \mathcal{X}_{\tilde{W}}$ be measurable functions.*

1. *$\tilde{M}$ is an exogenous pushforward of M through $\tilde{\Phi}$ if and only if M is an exogenous pullback of $\tilde{M}$ through $\tilde{\Phi}$.*

2. $\tilde{M}$ *is an exogenous pullback of* M *through* Φ *if and only if* M *is an exogenous pushforward of* M *through* Φ.

We can think of these two operations as *dual* to each other.

8.4. Parameterizing iSCMs using response functions

The following technique of "response functions" [BP94a] provides an example of an exogenous pushforward. It has been used—amongst others—to derive bounds on counterfactual probabilities and to obtain tests for valid instruments.

8.4.1. Example

First we will explain the idea using an example, postponing the general formal definition to Section 8.4.2.

Definition 8.4.1. *Let* $M = (J, V = \{v\}, W, \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_W, P, f)$ *be a simple iSCM with* $\mathcal{X}_J$ *discrete,* $\mathcal{X}_v$ *discrete, and* $\mathcal{X}_W$ *an arbitrary standard measurable space.*
Let $\tilde{W} = \{\tilde{w}\}$ *be a singleton and consider the space of (measurable) functions*[50]

$$\mathcal{X}_{\tilde{W}} := (\mathcal{X}_v)^{\mathcal{X}_J} = \{\phi : \mathcal{X}_J \to \mathcal{X}_v\}.$$

The iSCM M *induces a function* $\Phi : \mathcal{X}_W \to \mathcal{X}_{\tilde{W}}$ *that assigns to each exogenous random value* $x_W \in \mathcal{X}_W$ *the corresponding response function in* $\mathcal{X}_{\tilde{W}}$*, that is,* $\Phi(x_W)$ *is the function* $x_J \mapsto f_v(x_J, x_W)$*. The iSCM* $M_{\mathrm{repar}(\tilde{f},\Phi)} = \left(J, V = \{v\}, \tilde{W}, \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_{\tilde{W}}, \tilde{P}, \tilde{f}\right)$ *with as exogenous distribution the pushforward* $\tilde{P} = (\Phi)_*(P)$ *and causal mechanism* $\tilde{f}(x_J, x_{\tilde{W}}) = x_{\tilde{W}}(x_J)$ *(which just evaluates the response function* $x_{\tilde{W}}$ *in the input* x_J*) is called a* ***response variable parameterization*** *of* M.

We call it a parameterization because it preserves the important causal semantics:

Proposition 8.4.2. *The response variable parameterization of* M *in Definition 8.4.1 is an exogenous pushforward of* M *and is counterfactually equivalent to* M *w.r.t.* V.

Proof. Note that:

$$f(x_J, x_W) = \Phi(x_W)(x_J) = \tilde{f}(x_J, \Phi(x_W))$$

for all $x_J \in \mathcal{X}_J, x_W \in \mathcal{X}_W$. Because M and $\tilde{M}$ are both acyclic, they are both simple, and the claim now follows from Theorem 8.3.8, noting that Φ does not depend on X_J. □

Corollary 8.4.3. *Let* $M = (J, V, W, \mathcal{X}, P, f)$ *be a simple iSCM with discrete exogenous input space* $\mathcal{X}_J$*. Let* $v \in V$ *be an endogenous variable in* M *taking values in a discrete space* $\mathcal{X}_v$*. Then there exists an iSCM that is counterfactually equivalent to* M *w.r.t.* $\{v\}$ *that has just a single endogenous variable (*$V = \{v\}$ *with space* $\mathcal{X}_v$*) and exogenous random space* $(\mathcal{X}_v)^{\mathcal{X}_J}$.

[50] [BP94a] call these functions "response functions", and a random variable taking values in $\mathcal{X}_{\tilde{W}}$ a "response-function variable".

Proof. First marginalize out all endogenous variables except v, and then take the response variable parameterization of this marginal iSCM. It has the desired exogenous random space, and note that both operations preserve counterfactual equivalence w.r.t. $\{v\}$. □

8.4.2. General definition

The general definition of the response variable parameterization is more involved. Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM with discrete spaces $\mathcal{X}_J$ and $\mathcal{X}_V$ (but $\mathcal{X}_W$ can be an arbitrary standard measurable space).

First, let us consider the special case that $G(M)$ contains no bidirected edges. For each variable $v \in V$, consider $W_v := \mathrm{Pa}_{G^+(M)}(v) \cap W$, $V_v := \mathrm{Pa}_{G^+(M)}(v) \cap V$ and $J_v := \mathrm{Pa}_{G^+(M)}(v) \cap J$. Without loss of generality, we may assume that each causal mechanism f_v only depends on its parents (see Remark 6.9.5), and therefore write the structural equations as:

$$X_v = f_v(X_{J_v}, X_{V_v}, X_{W_v}).$$

For $v \in V$, we define the function space

$$\mathcal{X}_{\tilde{W}_v} := \mathcal{X}_v^{\mathcal{X}_{J_v} \times \mathcal{X}_{V_v}} = \{\mathcal{X}_{J_v} \times \mathcal{X}_{V_v} \to \mathcal{X}_v\}.$$

[BP94a] call these functions "response functions", and a random variable $X_{\tilde{W}_v}$ taking values in $\mathcal{X}_{\tilde{W}_v}$ a "response-function variable". We set $\tilde{W} := \{\tilde{W}_v : v \in V\}$ and $\mathcal{X}_{\tilde{W}} := \prod_{v \in V} \mathcal{X}_{\tilde{W}_v}$.

For each variable $v \in V$, the iSCM M induces a function $\tilde{\Phi}_v : \mathcal{X}_{W_v} \to \mathcal{X}_{\tilde{W}_v}$ that assigns to each exogenous state x_{W_v} the corresponding response function in $\mathcal{X}_{\tilde{W}_v}$, that is,

$$\tilde{\Phi}_v(x_{W_v}) := (x_{J_v}, x_{V_v}) \mapsto f_v(x_{J_v}, x_{V_v}, x_{W_v}).$$

Together, the $\tilde{\Phi}_v$ can be considered components of a map $\tilde{\Phi} : \mathcal{X}_W \to \mathcal{X}_{\tilde{W}}$, where $\tilde{\Phi}_v$ only depends on $\mathcal{X}_{W_v}$. We define the causal mechanism components (for $v \in V$)

$$\tilde{f}_v(x_J, x_V, x_{\tilde{W}}) := x_{\tilde{W}_v}(x_{J_v}, x_{V_v}),$$

which just evaluates the response function $x_{\tilde{W}_v} \in \mathcal{X}_{\tilde{W}_v}$ at (x_J, x_V). Note that by construction,

$$f(x_J, x_V, x_W) = \tilde{f}(x_J, x_V, \tilde{\Phi}(x_W))$$

for all $x \in \mathcal{X}$. We also define the pushforward exogenous distribution $\tilde{P} := (\tilde{\Phi})_* P$.

Definition 8.4.4. *The iSCM* $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})} = \left(J, V, \tilde{W}, \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_{\tilde{W}}, \tilde{P}, \tilde{f}\right)$ *constructed in this way is called the* ***response variable parameterization*** *of iSCM* $M = (J, V, W, \mathcal{X}, P, f)$ *for an iSCM* M *with discrete spaces* $\mathcal{X}_J$ *and* $\mathcal{X}_V$*, and no bidirected edges in* $G(M)$*.*

Now let us consider the general case. We will again construct a response function for every endogenous variable $v \in V$, but which response function is selected depends on all the exogenous parents in the *district* $\mathrm{Dist}^{G(M)}(v)$ of v. By effectively merging these exogenous parents we then end up with the required factorization of the reparameterized exogenous variables.

Let D be a partition of V into districts, i.e., $D = \{\mathrm{Dist}^{G(M)}(v) : v \in V\}$. We will denote by $d_v := \mathrm{Dist}^{G(M)}(v)$ the district of $v \in V$. For each variable $v \in V$, consider $W_v := \mathrm{Pa}_{G^+(M)}(v) \cap W$, $V_v := \mathrm{Pa}_{G^+(M)}(v) \cap V$ and $J_v := \mathrm{Pa}_{G^+(M)}(v) \cap J$.

For every $v \in V$, for every $x_{W_{d_v}} \in \mathcal{X}_{W_{d_v}}$ we define a "response function" $R_v(x_{W_{d_v}}) : \mathcal{X}_{J_v,V_v} \to \mathcal{X}_v$ that maps $(x_{J_v}, x_{V_v}) \mapsto f_v(x_{J_v}, x_{V_v}, x_{W_v})$. In other words, $R_v(x_{W_{d_v}}) \in \mathcal{X}_v^{\mathcal{X}_{J_v,V_v}} =: \mathcal{X}_{\tilde{W}_v}$. We define the spaces $\mathcal{X}_{\tilde{W}_d} := \prod_{v\in d} \mathcal{X}_{\tilde{W}_v}$, the Cartesian product of the response functions of the variables in district $d \in D$. We then define the mapping $\tilde{\Phi} : \mathcal{X}_W \to \prod_{d\in D} \mathcal{X}_{\tilde{W}_d}$ by

$$\tilde{\Phi}(x_W) = \big((R_v(x_{W_{d_v}}))_{v\in d}\big)_{d\in D}.$$

We define the causal mechanism components (for $v \in V$)

$$\tilde{f}_v(x_J, x_V, x_{\tilde{W}}) := (x_{\tilde{W}_{d_v}})_v(x_{J_v}, x_{V_v}),$$

which just evaluates the response function for variable v in district d at (x_J, x_V) (which by definition only depends on x_{J_v} and x_{V_v}). Note that by construction, for all $v \in V$,

$$f_v(x_J, x_V, x_W) = \tilde{f}_v(x_J, x_V, \tilde{\Phi}(x_W))$$

for all $x \in \mathcal{X}$. We define the pushforward exogenous distribution $\tilde{P} := (\tilde{\Phi})_* P$. We then have the required independence of the d-components of $\tilde{P}$, since $\tilde{\Phi}_d(X_W)$ only depends on X_{W_d}, and the $X_{W_d} = (X_v)_{v\in d}$ are independent in P.

Definition 8.4.5. *The iSCM $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})} = \left(J, V, \tilde{W}, \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_{\tilde{W}}, \tilde{P}, \tilde{f}\right)$ constructed in this way is called the* ***response variable parameterization*** *of iSCM $M = (J, V, W, \mathcal{X}, P, f)$ for an iSCM M with discrete spaces $\mathcal{X}_J$ and $\mathcal{X}_V$.*

Theorem 8.4.6. *Let $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$ be the response variable parameterization of simple iSCM M (which has discrete spaces $\mathcal{X}_J$ and $\mathcal{X}_V$). Then $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$ is simple and counterfactually equivalent to M w.r.t. V.*

Proof. By construction, $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$ has discrete space $\mathcal{X}_J$ and $\mathcal{X}_{\tilde{W}}$. Futhermore, $\tilde{\Phi}$ does not depend on X_J. If M is simple, then $M_{\mathrm{repar}(\tilde{f},\tilde{\Phi})}$ is simple by Proposition 8.3.9. Since both iSCMs are simple, $\tilde{\Phi}$ does not depend on X_J, the claim follows from Theorem 8.3.8. □

8.5. Bounding counterfactual probabilities

We have seen that unless one is willing to make very strong modeling assumptions, obtaining counterfactual probabilities from data can be impossible. In certain cases, though, it is possible to derive *bounds* on counterfactual probabilities from (intervened)

Markov kernels [BP94a]. One way to derive such bounds on counterfactuals exploits the response function parameterization.

As a motivation, consider a medical setting in which a patient may either be treated (or not) and a week later the patient is cured (or not). Suppose a patient participating in a randomized controlled trial was assigned to the control group and hence not treated ($\mathrm{do}(x = 0)$), and it turned out one week later that this patient was not cured ($Y^{\mathrm{do}(x=0)} = 0$). The patient now wonders "would I have been cured, had I been assigned to the treatment group?". By making use of the response-function parameterization, we can obtain a bound that does not depend on the specific parameters of the iSCM, yielding a "worst-case" lower bound and a "best-case" upper bound on the probability that the patient would then be cured.

Proposition 8.5.1 (Bounding counterfactual probabilities)**.** *For a simple iSCM M with a single binary exogenous input variable $X \in \{0,1\}$ and a binary endogenous variable $Y \in \{0,1\}$ (and perhaps additional endogenous variables, and with an arbitrary number of exogenous random variables taking values in arbitrary standard measurable spaces), the counterfactual probability*

$$P_{M^{\mathrm{twin}}}((Y')^{\mathrm{do}(x'=1)} = 1 \mid Y^{\mathrm{do}(x=0)} = 0)$$

(with $Y^{\mathrm{do}(x=0)}$ the factual outcome, and $(Y')^{\mathrm{do}(x'=1)}$ the counterfactual outcome) is bounded by

$$\frac{q_{0|0} - \min(q_{0|0}, q_{0|1})}{q_{0|0}} \leq P_{M^{\mathrm{twin}}}((Y')^{\mathrm{do}(x'=1)} = 1 | Y^{\mathrm{do}(x=0)} = 0) \leq \frac{\min(q_{0|0}, q_{1|1})}{q_{0|0}},$$

where $q_{y|x} := P_M(Y = y \mid \mathrm{do}(X = x))$.

Proof. Denote the 3-dimensional probability simplex by $\Delta_3 = \{r = (r_{00}, r_{01}, r_{10}, r_{11}) \in [0,1]^4 : r_{00} + r_{01} + r_{10} + r_{11} = 1\}$. We know from Corollary 8.4.3 that without loss of generality, we may assume that the iSCM has only a single binary endogenous variable Y and an exogenous random variable taking values in $\{0,1\}^{\{0,1\}}$. That iSCM must then lie in the family $\{M^\rho : \rho \in \Delta_3\}$, where the iSCM M^ρ with parameter ρ has binary exogenous input variable X, binary endogenous variable Y, a single latent exogenous random variable $W \in \{f_{00}, f_{01}, f_{10}, f_{11}\}$, exogenous distribution $P^\rho(W = f_w) = \rho_w$, and structural equation

$$Y^{\mathrm{do}(x)} = W(x),$$

where we defined response functions $f_{00}, f_{01}, f_{10}, f_{11} : \{0,1\} \to \{0,1\}$ by:

$$\begin{aligned} f_{00} &: 0 \mapsto 0, 1 \mapsto 0; \\ f_{01} &: 0 \mapsto 0, 1 \mapsto 1; \\ f_{10} &: 0 \mapsto 1, 1 \mapsto 0; \\ f_{11} &: 0 \mapsto 1, 1 \mapsto 1. \end{aligned}$$

We will derive a bound on the counterfactual probability

$$P_{(M^\rho)^{\mathrm{twin}}}((Y')^{\mathrm{do}(x'=1)} = 1 \mid Y^{\mathrm{do}(x=0)} = 0).$$

We first update the distribution of W with the observed outcome:

$$P_{(M^\rho)^{\text{twin}}}(W = w | Y^{\text{do}(x=0)} = 0) = \begin{cases} \frac{\rho_{00}}{\rho_{00}+\rho_{01}} & w = f_{00}, \\ \frac{\rho_{01}}{\rho_{00}+\rho_{01}} & w = f_{01}, \\ 0 & w = f_{10}, \\ 0 & w = f_{11}. \end{cases}$$

Because of the counterfactual equivalence of M^ρ and M w.r.t. Y, we have that $q_{y|x} := P_M(Y = y \,|\, \text{do}(X = x)) = P_{(M^\rho)^{\text{twin}}}(Y = y \,|\, \text{do}(X = x))$. This Markov kernel is given explicitly by:

x	y	$q_{y\|x}$
0	0	$\rho_{00} + \rho_{01}$
0	1	$\rho_{10} + \rho_{11}$
1	0	$\rho_{00} + \rho_{10}$
1	1	$\rho_{01} + \rho_{11}$

For our particular counterfactual probability of interest, we have the equality

$$P_{(M^\rho)^{\text{twin}}}((Y')^{\text{do}(x'=1)} = 1 | Y^{\text{do}(x=0)} = 0) = \frac{\rho_{01}}{q_{0|0}}.$$

From the non-negativity of the ρ's and the table above, we can derive the bound

$$q_{0|0} - \min(q_{0|0}, q_{0|1}) \leq \rho_{01} \leq \min(q_{0|0}, q_{1|1})$$

and hence

$$\frac{q_{0|0} - \min(q_{0|0}, q_{0|1})}{q_{0|0}} \leq P_{(M^\rho)^{\text{twin}}}((Y')^{\text{do}(x'=1)} = 1 | Y^{\text{do}(x=0)} = 0) \leq \frac{\min(q_{0|0}, q_{1|1})}{q_{0|0}}.$$

Since M^ρ is counterfactually equivalent to the original iSCM M w.r.t. Y, the same bound also holds for M instead of M^ρ. □

As an illustration, suppose that $q_{0|0} \approx 1$ and $q_{1|1} \approx 1$. Then the bound tells us that $P_{M^{\text{twin}}}((Y')^{\text{do}(x'=1)} = 1 | Y^{\text{do}(x=0)} = 0) \approx 1$ as well. Thus, for almost-deterministic relations, we can tightly bound this counterfactual probability.

9. Causal Relations and Confounding

In this chapter, we will formalize various elementary causal relations. Our treatment extends conventional notions as we explicitly allow for exogenous input nodes to represent hard interventions with unspecified intervention values. For each causal relation, we give a graphical and several iSCM versions of the notion, and show how these are related. Typically, the absence of a certain causal relation in the graph implies its absence in the iSCM, but not vice versa. We will also connect these versions to certain patterns in a set of observable and interventional Markov kernels, and to various formulations in terms of potential outcomes and counterfactuals.

9.1. Faithfulness

Before we start, we introduce a certain "genericity" notion called *faithfulness*. It is the converse statement of the global Markov property: loosely speaking it says that every conditional independence in the Markov kernel of a causal model must correspond with a d-separation or σ-separation in the causal graph corresponding to the model. Originally, this notion was only defined for the d-separation Markov property (applicable to acyclic models and certain subclasses of cyclic iSCMs) and referred to as "faithfulness". To avoid confusion, we explicitly distinguish two versions: σ-faithfulness for simple iSCMs (corresponding to Corollary 7.2.1), and d-faithfulness for acyclic iSCMs and causal Bayesian networks (corresponding to Theorem 4.2.1).

Definition 9.1.1. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM with causal graph $G(M)$ and Markov kernel $P_M(X_V \,|\, \mathrm{do}(X_J))$. M is called **σ-faithful** if for all $A, B, C \subseteq J \cup V$ (not necessarily disjoint):*

$$A \underset{G(M)}{\overset{\sigma}{\perp}} B \,|\, C \iff X_A \underset{P_M(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C. \tag{47}$$

*M is called **d-faithful** if for all $A, B, C \subseteq J \cup V$ (not necessarily disjoint):*

$$A \underset{G(M)}{\overset{d}{\perp}} B \,|\, C \iff X_A \underset{P_M(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C. \tag{48}$$

*For a subset $O \subseteq V$, we say that M is **σ-faithful w.r.t.** O if (47) holds for all (not necessarily disjoint) $A, B, C \subseteq J \cup O$, and we define d-**faithful w.r.t.** O analogously.*

In words, a simple iSCM is called σ-faithful (d-faithful) if each conditional independence in the induced Markov kernel is due to a σ-separation (d-separation) in its causal graph. Note that σ-faithfulness is a stronger assumption than d-faithfulness (as σ-separation implies d-separation).

Remark 9.1.2. *These notions behave properly under marginalization: a simple iSCM M is σ/d-faithful w.r.t. $O \subseteq V$ if and only if its marginalizations M_O are σ/d-faithful.*

Faithfulness may fail for various reasons:

- Deterministic relationships may lead to additional conditional independences, but are not exploited by the Markov property;
- Effects may cancel out;
- If cycles are present, and (i) all variables are discrete, or (ii) interactions are linear, σ-faithfulness may fail;
- If cycles are present, and the system is "perfectly adapting".

An example of a deterministic relationship leading to a faithfulness violation is the following.

Example 9.1.3. *Take an SCM M with three endogenous variables X, Y, Z and two exogenous random variables U, W, with structural equations*

$$X = 5, \qquad Y = X + U, \qquad Z = X + W.$$

Then $Y \not\perp^{\sigma}_{G(M)} Z$ and $Y \not\perp^{d}_{G(M)} Z$, but $Y \perp\!\!\!\perp_{P_M} Z$. This simple (even acyclic) SCM is neither σ-faithful nor d-faithful, due to X being constant.

The next example illustrates how canceling effects may lead to a faithfulness violation.

Example 9.1.4. *Take an SCM M with three endogenous variables X, Y, Z and three exogenous random variables W_X, W_Y, W_Z, with structural equations*

$$X = W_X, \qquad Y = X + W_Y, \qquad Z = Y - X + W_Z.$$

Then $X \not\perp^{\sigma}_{G(M)} Z$ and $X \not\perp^{d}_{G(M)} Z$, but $X \perp\!\!\!\perp_{P_M} Z$. This simple (even acyclic) SCM is neither σ-faithful nor d-faithful, due to cancellation of the direct causal influences of X on Z with the indirect causal influence of X via Y on Z.

One can show that in certain special cases, the global Markov property in terms of d-separation even holds for cyclic simple iSCMs.

Proposition 9.1.5. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM with causal graph $G(M)$ and Markov kernel $P_M(X_V \,|\, \mathrm{do}(X_J))$. If $J = \emptyset$ and one of the three conditions applies:*

1. *M is acyclic, or*
2. *all spaces $\mathcal{X}_v$ with $v \in V$ are discrete, or*
3. *M is essentially linear over the field $\mathbb{R}$ of the following restricted form: the endogenous spaces are $\mathcal{X}_v = \mathbb{R}^{d_v}$ for $d_v \in \mathbb{N}$ for $v \in V$, the causal mechanisms are of the form*
$$f_v(x_W, x_V) = \sum_{u \in V} B_{vu} x_u + c_v(x_W) \qquad M\text{-a.s.},$$
for $v \in V$, where $B_{vu} : \mathcal{X}_u \to \mathcal{X}_v$ is a linear mapping for $u, v \in V$ and $c_v : \mathcal{X}_W \to \mathcal{X}_v$ is a measurable function for $v \in V$,

then for all $A, B, C \subseteq J \cup V$ (not necessarily disjoint):

$$A \underset{G(M)}{\overset{d}{\perp}} B \,|\, C \implies X_A \underset{P_M(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C.$$

Proof. The proofs are essentially given in [FM17].

By Remark 6.9.5, we can assume without loss of generality that for each $v \in V$, f_v is constant in $x_{(J \cup W \cup V) \setminus \mathrm{Pa}^{G^+(M)}(v)}$. We give a sketch for each of the cases.

1. If M is acyclic, then $G^+(M)$ is acyclic. The claim now follows from the global Markov property Theorem 4.2.1 for M interpreted as a causal Bayesian network as in the proof of Proposition 6.11.4 point i) (with deterministic Markov kernels for the endogenous variables, and purely probabilistic Markov kernels for the exogenous random variables).

2. If all spaces $\mathcal{X}_v$ with $v \in V$ are discrete, then one can apply Theorem 3.8.12, Remark 3.7.2, Theorem 3.6.6 and Theorem 3.5.2 in [FM17].

3. This extends in [FM17, Example 3.8.17] in two ways. First, that example assumes $d_v = 1$ for all $v \in V$, but its proof immediately generalizes to arbitrary $(d_v)_{v \in V}$. Second, that example assumes that $c(x_W)$ has a density w.r.t. Lebesgue measure on $\mathbb{R}^d$ with $d = \sum_{v \in V} d_v$. This extension proceeds in the following way.

   For $\epsilon > 0$, construct a perturbed iSCM $M^{(\epsilon)}$ out of M that has additional exogenous random variables $\tilde{w}_v$ for all $v \in V$, with corresponding spaces $\mathcal{X}_{\tilde{w}_v} := \mathcal{X}_v = \mathbb{R}^{d_v}$ and (independent) exogenous distributions $X_{\tilde{w}_v} \sim \mathcal{N}(0, \epsilon^2 I_v)$ with I_v the $d_v \times d_v$ identity matrix. By [BW16, Corollary 2.2], the distribution of $\tilde{c}(x_W, x_{\tilde{W}}) := c(X_W) + X_{\tilde{W}}$ has a density w.r.t. the Lebesgue measure on $\mathbb{R}^d$. Hence, by (the extension of) [FM17, Example 3.8.17], the d-separation Markov property holds in $M^{(\epsilon)}$. In other words, for all $A, B, C \subseteq J \cup V$ (not necessarily disjoint):

$$A \underset{G(M^{(\epsilon)})}{\overset{d}{\perp}} B \,|\, C \implies X_A \underset{P_{M^{(\epsilon)}}(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C.$$

   As $\epsilon \to 0$, $P_{M^{(\epsilon)}}(\tilde{c}(X_W, X_{\tilde{W}})) \to P_M(c(X_W))$ in total variation distance.[51] Since M is simple, the matrix $(I - B)$ is invertible and hence

$$P_M(X_V) = ((I - B)^{-1})_* P_M(c(X_W))$$

   and, consequently, also $M^{(\epsilon)}$ is simple and

$$P_{M^{(\epsilon)}}(X_V) = ((I - B)^{-1})_* P_{M^{(\epsilon)}}(\tilde{c}(X_W, X_{\tilde{W}})).$$

   A measurable transformation can only decrease total variation distance. Indeed, let X_1, X_2 be two random variables taking values in a measurable space $\mathcal{X}$, with

[51] The *total variation distance* of the distributions of two random variables X_1, X_2 taking values in the same space $\mathcal{X}$ is defined as $d_{TV}(P(X_1), P(X_2)) := \sup_{A \in \mathcal{B}_{\mathcal{X}}} |P(X_1 \in A) - P(X_2 \in A)|$.

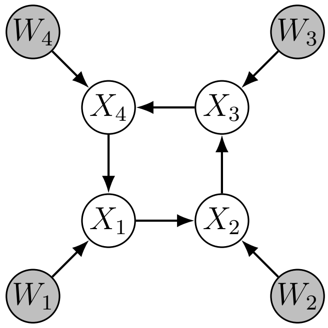


Figure 24: Graph $G^+(M)$ of SCM M of Example 9.1.6. $X_1 \perp^d_{G^+(M)} X_3 \mid X_2, X_4$, but $X_1 \not\perp^\sigma_{G^+(M)} X_3 \mid X_2, X_4$. M is d-faithful, but not σ-faithful.

distributions $P(X_1)$, $P(X_2)$, respectively. Define $Y_1 := f(X_1)$ and $Y_2 := f(X_2)$ for some $\mathcal{B}_\mathcal{X}$-$\mathcal{B}_\mathcal{Y}$-measurable function $f : \mathcal{X} \to \mathcal{Y}$. Then

$$\begin{aligned} d_{TV}(P(Y_1), P(Y_2)) &= \sup_{B \in \mathcal{B}_\mathcal{Y}} |P(Y_1 \in B) - P(Y_2 \in B)| \\ &= \sup_{B \in \mathcal{B}_\mathcal{Y}} |P(X_1 \in f^{-1}(B)) - P(X_2 \in f^{-1}(B))| \\ &\leq \sup_{A \in \mathcal{B}_\mathcal{X}} |P(X_1 \in A) - P(X_2 \in A)| \\ &= d_{TV}(P(X_1), P(X_2)). \end{aligned}$$

Hence, $P_{M^{(\epsilon)}}(X_V)$ converges to $P_M(X_V)$ in total variation distance as $\epsilon \to 0$. Since conditional independences are preserved under convergence in total variation distance [Lau24, Theorem 1] [CE24, Lemma A.1], and $G(M^{(\epsilon)}) = G(M)$, the claim follows.

□

The following gives an example of a cyclic simple SCM that is d-faithful but not σ-faithful.

Example 9.1.6. *Consider an SCM M with four endogenous variables X_1, X_2, X_3, X_4, four exogenous random variables W_1, W_2, W_3, W_4, with structural equations*

$$\begin{aligned} X_1 &= X_4 + W_1 \\ X_2 &= X_1 + W_2 \\ X_3 &= X_2 + W_3 \\ X_4 &= \tfrac{1}{2}(X_3 + W_4). \end{aligned}$$

We can solve for X in terms of W:

$$\begin{aligned} X_1 &= 2W_1 + W_2 + W_3 + W_4 \\ X_2 &= 2W_1 + 2W_2 + W_3 + W_4 \\ X_3 &= 2W_1 + 2W_2 + 2W_3 + W_4 \\ X_4 &= W_1 + W_2 + W_3 + W_4 \end{aligned}$$

No matter what the distributions $P(W_i)$ are, we have $X_1 \perp\!\!\!\perp X_3 \mid X_2, X_4$ because $1 \perp^d 3 \mid 2, 4$ in $G^+(M)$ (see Figure 24). However, $1 \not\perp^\sigma 3 \mid 2, 4$ in $G^+(M)$. Therefore, this SCM is not σ-faithful. One can check explicitly that it is d-faithful if each $P(W_i)$ is non-degenerate.

9.2. Causal Relations

We are now ready to formalize the notion of "a causes b" (finally!) of Definition 1.2.3. We will formulate this notion at different levels of the causal hierarchy. At the highest level, we have the graphical notion. Then comes the counterfactual level (or potential outcome level), and finally the interventional level.[52]

9.2.1. Causal relations (graphical notion)

We will start at the graphical level. We typically give the following causal interpretation to a CDMG.

Definition 9.2.1. *Let G be a CDMG with input nodes J and output nodes V. Let $a \in J \cup V$ and $b \in V$. If $a \notin \mathrm{Anc}^G(b)$ then we say that a **does not cause** b **according to** G; otherwise, we say that a **causes** b **according to** G.*

Remark 9.2.2. *Marginalization of a graph preserves its causal relations (see also Remark 3.2.19). In addition, we have the equivalences:*

$$\begin{aligned} & a \text{ causes } b \text{ according to } G \\ \iff\quad & a \longrightarrow b \text{ is present in } G^{\setminus(V\setminus\{a,b\})} \\ \iff\quad & a \text{ causes } b \text{ according to } G_{\mathrm{do}(a)} \\ \iff\quad & I_a \text{ causes } b \text{ according to } G_{\mathrm{do}(I_a)}. \end{aligned}$$

There is one more useful equivalence, which expresses the graphical notion of causation in terms of a separation statement.

Proposition 9.2.3. *Let G be a CDMG with input nodes J and output nodes V. Let $b \in V$. For $a \in J \cup V$:*

$$a \notin \mathrm{Anc}^G(b) \iff b \underset{G_{\mathrm{do}(I_a)}}{\overset{\sigma}{\perp}} I_a \mid J \setminus \{I_a\}. \tag{49}$$

Proof. Suppose that there exists a walk in $G_{\mathrm{do}(I_a)}$ between b and $I_a \cup J$ that is σ-open given $J\setminus\{I_a\}$. Then there exists such a walk without colliders (Proposition 3.3.6 and the observation that no node in J can be a collider on a walk). Such a walk cannot contain a node from $J \setminus \{I_a\}$, since that would either be an end node (σ-blocking the walk) or a non-collider node pointing only to nodes in another strongly connected component (σ-blocking the walk). Therefore it must be of the form $I_a \longrightarrow a \longrightarrow \cdots b$. Since it cannot contain a collider, it must be a directed walk, and hence $I_a \in \mathrm{Anc}^{G_{\mathrm{do}(I_a)}}(b)$. Vice versa, any directed walk from I_a to b in $G_{\mathrm{do}(I_a)}$ is σ-open given $J \setminus \{I_a\}$, as it cannot contain a node in $J \setminus \{I_a\}$. $\square$

[52]Some refer to the counterfactual notion of causation as "individual" or "token" causation, and to the interventional notion of causation as "population" or "type" causation.

Remark 9.2.4. *We can rewrite the r.h.s. of (49), distinguishing the cases $a \in J$ and $a \in V$, as follows:*

$$b \underset{G_{\mathrm{do}(I_a)}}{\overset{\sigma}{\perp}} I_a \mid J \setminus \{I_a\} \iff \begin{cases} b \perp_G^{\sigma} a \mid J \setminus \{a\} & a \in J \\ b \perp_{G_{\mathrm{do}(I_a)}}^{\sigma} I_a \mid J & a \in V. \end{cases}$$

9.2.2. Causal relations (interventional notion)

The most common notion of causation is the interventional one.

Definition 9.2.5. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let $a \in J \cup V$ and $b \in V$. If*

$$X_b \underset{P_{M_{\mathrm{do}(I_a)}}}{\perp\!\!\!\perp} X_{I_a} \mid X_{J \setminus \{I_a\}}, \tag{50}$$

*we say that a **is not a rung-2 cause of** b **according to** M. Otherwise, we say that a **is a rung-2 cause of** b **according to** M.*

Remark 9.2.6. *We can rewrite (50), distinguishing the cases $a \in J$ and $a \in V$, as follows:*

$$X_b \underset{P_{M_{\mathrm{do}(I_a)}}}{\perp\!\!\!\perp} X_{I_a} \mid X_{J \setminus \{I_a\}} \iff \begin{cases} X_b \perp\!\!\!\perp_{P_M} X_a \mid X_{J \setminus \{a\}} & a \in J, \\ X_b \perp\!\!\!\perp_{P_{M_{\mathrm{do}(I_a)}}} X_{I_a} \mid X_J & a \in V. \end{cases}$$

Alternatively, we can write (50) as:

$$P_M(X_b \mid \mathrm{do}(X_{J \cup \{a\}})) = P_M(X_b \mid \mathrm{do}(X_{J \setminus \{a\}})),$$

or, distinguishing the two cases, as:

$$\begin{cases} P_M(X_b \mid \mathrm{do}(X_J)) = P_M(X_b \mid \mathrm{do}(X_{J \setminus \{a\}}), \cancel{\mathrm{do}(X_a)}) & a \in J, \\ P_M(X_b \mid \mathrm{do}(X_J), \mathrm{do}(X_a)) = P_M(X_b \mid \mathrm{do}(X_J)) & a \in V. \end{cases}$$

Remark 9.2.7. *Marginalization of an iSCM preserves the rung-2 causal relations between the remaining variables.*

Remark 9.2.8. *Rung-2 causal relations are invariant under interventional equivalence of iSCMs. More precisely: a is a rung-2 cause of b according to M implies that a is a rung-2 cause of b according to $\tilde{M}$ if M is interventionally equivalent to $\tilde{M}$ w.r.t. $\{a, b\} \cap V$ (remember that Proposition 6.7.8 states that if M and $\tilde{M}$ are interventionally equivalent w.r.t. $\{a, b\} \cap V \subseteq V \cap \tilde{V}$, then so are $M_{\mathrm{do}(I_a)}$ and $\tilde{M}_{\mathrm{do}(I_a)}$ if $a \in (V \cap \tilde{V}) \cup (J \cap \tilde{J})$).*

Remark 9.2.6 shows that the two cases $a \in V$ and $a \in J$ mean something subtly different, and one may wonder why we chose to work with $M_{\mathrm{do}(I_a)}$ rather than $M_{\mathrm{do}(a)}$ (which would treat the two cases in the same way). The reason is that the observable Markov kernel may provide additional information beyond that provided by the interventional Markov kernels, as the following example shows.

Example 9.2.9. *Consider the iSCM M with endogenous variables $X_a, X_b \in \{-1,+1\}$, exogenous random variable $X_w \in \{-1,1\}$, and structural equations:*

$$X_a = X_w,$$
$$X_b = X_a X_w,$$

where $X_w \sim \mathcal{U}(\{-1,1\})$. Note that $P_M(X_b = 1 \,|\, \mathrm{do}(X_a = x_a)) = \frac{1}{2}$ for $x_a \in \{-1,1\}$. However, $P_M(X_b = 1) = 1$. Hence, a is a rung-2 cause of b according to M. So we cannot detect that a causes b from the interventional Markov kernel $P_M(X_b \,|\, \mathrm{do}(X_a))$ only, but we can if we additionally compare with the observational distribution $P_M(X_b)$.

Consider now the iSCM $M_{\mathrm{do}(a)}$, with endogenous variable $X_b \in \{-1,+1\}$, exogenous random variable $X_w \in \{-1,+1\}$, exogenous input variable $X_a \in \{-1,+1\}$, and structural equation:

$$X_b = X_a X_w.$$

According to $M_{\mathrm{do}(a)}$, a is not a rung-2 cause of b.

We conclude that a can be a rung-2 cause of b according to some iSCM M, while a is not a rung-2 cause of b according to $M_{\mathrm{do}(a)}$ (which might be a little surprising at first sight).[53]

9.2.3. Causal relations (counterfactual notion)

A more 'refined' notion of causation (but also harder to detect in practice) can be obtained by considering counterfactuals.

Definition 9.2.10. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let $a \in J \cup V$ and $b \in V$. We say that a **is not a rung-3 cause of** b **according to** M if:*

$$\begin{aligned}&\forall x_{J\setminus\{I_a\}} \in \mathcal{X}_{J\setminus\{I_a\}}, \forall x_{I_a}, x_{I_{a'}} \in \mathcal{X}_{I_a} : \\ &P_{(M_{\mathrm{do}(X_{J\setminus\{I_a\}}=x_{J\setminus\{I_a\}}),\mathrm{do}(I_a)})^{\mathrm{twin}}}(X_b = X_{b'} \,|\, \mathrm{do}(X_{I_a} = x_{I_a}), \mathrm{do}(X_{I_{a'}} = x_{I_{a'}})) = 1.\end{aligned} \tag{51}$$

*Otherwise, we say that a **is a rung-3 cause of** b **according to** M.*

Remark 9.2.11. *We can rewrite (51) for $a \in V$ as:*

$$\forall x_J \in \mathcal{X}_J, \forall x_a \in \mathcal{X}_a : \quad P_{((M_{\mathrm{do}(X_J=x_J)})^{\mathrm{twin}})_{\mathrm{do}(a)}}(X_b = X_{b'} \,|\, \mathrm{do}(X_a = x_a)) \neq 1;$$

and for $a \in J$ as:

$$\forall x_{J\setminus\{a\}} \in \mathcal{X}_{J\setminus\{a\}}, \forall x_a, x'_a \in \mathcal{X}_a : \quad P_{(M_{\mathrm{do}(X_{J\setminus\{a\}}=x_{J\setminus\{a\}})})^{\mathrm{twin}}}(X_b = X_{b'} \,|\, \mathrm{do}(X_a = x_a), \mathrm{do}(X_{a'} = x'_a)) \neq 1.$$

Remark 9.2.12. *Marginalization of an iSCM preserves the rung-3 causal relations of the remaining variables.*

[53] This example also shows that the claim in [PJS17, Proposition 6.13] is incorrect.

Remark 9.2.13. *Rung-3 causal relations are invariant under counterfactual equivalence of iSCMs. More precisely: a is a rung-3 cause of b according to M implies that a is a rung-3 cause of b according to $\tilde{M}$ if M is counterfactually equivalent to $\tilde{M}$ w.r.t. $\{a,b\} \cap V$ (remember that Proposition 8.2.6 states that if M and $\tilde{M}$ are counterfactually equivalent w.r.t. $\{a,b\} \cap V \subseteq V \cap \tilde{V}$, then so are $M_{\mathrm{do}(I_a)}$ and $\tilde{M}_{\mathrm{do}(I_a)}$ if $a \in (V \cap \tilde{V}) \cup (J \cap \tilde{J})$).*

The following example shows that the presence of a rung-3 causal relation does not necessarily imply the presence of a rung-2 causal relation. Intuitively, the causal effect can go unnoticed if it averages out.

Example 9.2.14. *Consider the iSCM M with exogenous input variable $X_a \in \{-1,+1\}$, endogenous variable $X_b \in \{-1,+1\}$, exogenous random variable $X_w \in \{-1,1\}$, and structural equation:*

$$X_b = X_a X_w,$$

where $X_w \sim \mathcal{U}(\{-1,1\})$. Note that a is a rung-3 cause of b according to M, but a is not a rung-2 cause of b according to M. Indeed, due to the uniform distribution of X_w the dependence of b on a averages out.

So, for example, a drug can have a beneficial effect on all males and an adversary effect on all females, and still have *no effect* on the entire population (if the positive effect exactly cancels out the negative effect on average).

The next example shows that the presence of a causal relation according to the causal graph does not necessarily imply the presence of a rung-3 causal relation. Intuitively, different causal paths can cancel out.

Example 9.2.15. *Consider the iSCM M with exogenous input variable $X_a \in \mathbb{R}$, endogenous variables $X_c, X_d, X_b \in \mathbb{R}$, exogenous random variable $X_w \in \mathbb{R}$, and structural equations*

$$\begin{aligned}
X_c &= X_a, \\
X_d &= -X_a \\
X_b &= X_c + X_d + X_w
\end{aligned}$$

with $X_w \sim \mathcal{N}(0,1)$. Then a causes b according to $G(M)$, but a is not a rung-3 cause of b according to M.

9.2.4. Causal relations (potential outcome notion)

Yet another possible notion (at first sight) of causation can be given in terms of potential outcomes.

Definition 9.2.16. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let $a \in J \cup V$ and $b \in V$. Let g be a solution function of $M_{\mathrm{do}(I_a)}$ and let $X_W \sim P(X_W)$ be a random variable. We say that a* ***is not a potential-outcome cause of*** *b* ***according to*** *M if:*

$$\forall x_{J\setminus\{I_a\}} \in \mathcal{X}_{J\setminus\{I_a\}}, \forall x_{I_a}, x'_{I_a} \in \mathcal{X}_{I_a} : \qquad g_b(x_{I_a}, x_{J\setminus\{I_a\}}, X_W) = g_b(x'_{I_a}, x_{J\setminus\{I_a\}}, X_W) \text{ a.s..}$$

Otherwise, we say that a ***is a potential-outcome cause of*** *b* ***according to*** *M.*

Using the notation of potential outcomes, we can also write:

$$\forall x_{J\setminus\{I_a\}} \in \mathcal{X}_{J\setminus\{I_a\}}, \forall x_{I_a}, x'_{I_a} \in \mathcal{X}_{I_a} : \qquad X_b^{\mathrm{do}(x_{I_a}, x_{J\setminus\{I_a\}})} = X_b^{\mathrm{do}(x'_{I_a}, x_{J\setminus\{I_a\}})} \text{ a.s.}$$

Here it is very important to keep in mind that we must use the same X_W to induce $X_b^{\mathrm{do}(x_{I_a}, x_{J\setminus\{I_a\}})} := g_b(x_{I_a}, x_{J\setminus\{I_a\}}, X_W)$ and $X_b^{\mathrm{do}(x'_{I_a}, x_{J\setminus\{I_a\}})} := g_b(x'_{I_a}, x_{J\setminus\{I_a\}}, X_W)$.

Remark 9.2.17. *Note that the choice of W does not matter for the definition; any random variable that is distributed according to the exogenous distribution of M can be used. Furthermore, the choice of g does not matter for the definition by Lemma 6.3.13; any solution function of $M_{\mathrm{do}(I_a)}$ can be used.*

At second sight, it turns out that the potential-outcome notion and the counterfactual notion coincide.

Proposition 9.2.18. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let $a \in J \cup V$ and $b \in V$. Then a is a rung-3 cause of b if and only if a is a potential-outcome cause of b.*

Proof. This is a matter of writing out the definitions. □

The potential-outcome notion has the advantage that it is considerably simpler to formulate: it leads to shorter expressions than the ones involving the twinning operation.

9.2.5. Hierarchy of causal relations

The following result shows the logical relations between various notions of causation. It is somewhat reminiscent of Pearl's causal hierarchy.

Theorem 9.2.19. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM with causal graph $G(M)$. Let $a \in V \cup J$ and $b \in V$. Consider the following statements:*

(i) a does not cause b according to $G(M)$;

(ii) a is not a rung-3 cause of b according to M;

(ii) a is not a rung-2 cause of b according to M.

Then (i) $\implies$ (ii) $\implies$ (iii). If $M_{\mathrm{do}(I_a)}$ is σ-faithful, also (iii) $\implies$ (i).

Proof. Pick an arbitrary $x^0_{I_a} \in \mathcal{X}_{I_a}$ (for example, $x^0_{I_a} = \star$ for $a \in V$). Write "$\forall\!\!\!\forall x_w \in \mathcal{X}_W$" as a shorthand of "for P-almost all $x_w \in \mathcal{X}_W$". Consider the properties:[54]

(a) $a \notin \mathrm{Anc}^{G(M)}(b)$;

(b) there exists a solution function g of $M_{\mathrm{do}(I_a)}$ such that: $\forall x_{J\setminus\{I_a\}} \in \mathcal{X}_{J\setminus\{I_a\}}$, $\forall\!\!\!\forall x_W \in \mathcal{X}_W$, $\forall x_{I_a} \in \mathcal{X}_{I_a}$: $g_b(x_{I_a}, x_{J\setminus\{I_a\}}, x_W) = g_b(x^0_{I_a}, x_{J\setminus\{I_a\}}, x_W)$;

[54]Note that the ordering of the quantifiers matters! In particular, $\forall\!\!\!\forall x_W \forall x_V \implies \forall x_V \forall\!\!\!\forall x_W$, whereas the converse doesn't hold in general (but does hold if $\mathcal{X}_V$ is discrete).

(c) for all solution functions g of $M_{\mathrm{do}(I_a)}$:
$\forall x_{J\setminus\{I_a\}} \in \mathcal{X}_{J\setminus\{I_a\}}, \forall x_{I_a} \in \mathcal{X}_{I_a}, \not\forall x_W \in \mathcal{X}_W$: $g_b(x_{I_a}, x_{J\setminus\{I_a\}}, x_W) = g_b(x^0_{I_a}, x_{J\setminus\{I_a\}}, x_W)$;

(d) for all solution functions g of $M_{\mathrm{do}(I_a)}$:
$\forall x_{J\setminus\{I_a\}} \in \mathcal{X}_{J\setminus\{I_a\}}, \forall x_{I_a} \in \mathcal{X}_{I_a}$: $P_M(g_b(x_{I_a}, x_{J\setminus\{I_a\}}, X_W)) = P_M(g_b(x^0_{I_a}, x_{J\setminus\{I_a\}}, X_W))$.

We will show that (a) $\implies$ (b) $\implies$ (c) $\implies$ (d), and that (d) $\implies$ (a) if $M_{\mathrm{do}(I_a)}$ is σ-faithful. The claim then follows by noting that (i) is equivalent to (a), (ii) to (c), and (iii) to (d).

We show:
(a) $\implies$ (b): Noting that $a \notin \mathrm{Anc}^{G(M)}(b) \iff a \notin \mathrm{Anc}^{G^+(M)}(b) \iff I_a \notin \mathrm{Anc}^{G^+(M_{\mathrm{do}(I_a)})}(b)$, Lemma 9.2.20 applied to $M_{\mathrm{do}(I_a)}$ provides us with a solution function g of $M_{\mathrm{do}(I_a)}$ for which g_b is essentially constant in x_{I_a}.

(b) $\implies$ (c). For the solution function g that satisfies (b), it follows (by swapping quantifiers) that: $\forall x_{J\setminus\{I_a\}} \in \mathcal{X}_{J\setminus\{I_a\}}, \forall x_{I_a} \in \mathcal{X}_{I_a}, \not\forall x_W \in \mathcal{X}_W$: $g_b(x_{I_a}, x_{J\setminus\{I_a\}}, x_W) = g_b(x^0_{I_a}, x_{J\setminus\{I_a\}}, x_W)$. In other words,

$$g_b(x_{I_a}, x_{J\setminus\{I_a\}}, x_W) = g_b(x^0_{I_a}, x_{J\setminus\{I_a\}}, x_W)\ M\text{-a.s..}$$

Because the solution function of M is essentially unique, this property actually holds for any solution function of M.

(c) $\implies$ (d) is trivial: almost-sure equality implies equality in distribution.

(d) $\implies$ (a) follows from the definition of σ-faithfulness in combination with Proposition 9.2.3. $\square$

Lemma 9.2.20. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Then there exists a solution function g of M such that for all $b \in V$, $g_b : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_b$ is constant in $(J \cup W) \setminus \mathrm{Anc}^{G^+(M)}(b)$:*

$$\forall x \in \mathcal{X} : \qquad g_b(x_J, x_W) = g_b(x_{(J\cup W)\cap \mathrm{Anc}^{G^+(M)}(b)}, x_{\cancel{\mathrm{Anc}^{G^+(M)}(b)}}).$$

Proof. Without loss of generality, we may assume that each causal mechanism f_v only depends on its parents (see Remark 6.9.5), and thus consider $f_v : \mathcal{X} \to \mathcal{X}_v$ as a function $f_v : \mathcal{X}_{\mathrm{Pa}^{G^+(M)}(v)} \to \mathcal{X}_v$ instead, for all $v \in V$.

We construct g as follows. Let $C \subseteq V$ be a strongly connected component of $G(M)$. We can pick a partial solution function $g^{[C]}$ of M w.r.t. C that only depends on $\mathrm{Pa}^{G^+(M)}(C) \setminus C$. Indeed, any variable that appears in a causal mechanism f_c for $c \in C$ must be in $\mathrm{Pa}^{G^+(M)}(c)$, that is, $g^{[C]}$ solves the equations

$$x_C = f_C(x_{\mathrm{Pa}^{G^+(M)}(C)})$$

for x_C in terms of $x_{\mathrm{Pa}^{G^+(M)}(C)\setminus C}$. Since the equations have a unique solution up to an M-null set, we can pick $g^{[C]} : \mathcal{X}_{\mathrm{Pa}^{G^+(M)}(C)\setminus C} \to \mathcal{X}_C$ to be that unique solution. This function is constant in $x_{J\cup W\setminus(\mathrm{Pa}^{G^+(M)}(C)\setminus C)}$ (up to an M-null set), and hence we can extend it to be constant in $x_{J\cup W\setminus(\mathrm{Pa}^{G^+(M)}(C)\setminus C)}$ everywhere. Construct such $g^{[C]}$ for each strongly connected component $C \subseteq V$ of $G(M)$ and consider the corresponding acyclification M^{acy} (cf. Definition 7.1.1); this can be considered a "structurally minimal acyclification of M". We then obtain a global solution function g of M^{acy} by recursive composition of the partial solution functions $g^{[C]}$ (cf. the proof of Proposition 7.1.2). It has the property that for all $v \in V$, g_v is constant in the non-ancestors of v according to $G^+(M^{\mathrm{acy}})$. But this g is also a solution function of M. It has the desired property, since $\mathrm{Anc}^{G^+(M^{\mathrm{acy}})}(v) \subseteq \mathrm{Anc}^{G^+(M)}(v)$ for all $v \in V$. □

We have already seen that the reverse implications in Theorem 9.2.19 may not hold. Indeed, Example 9.2.15 exhibits a graphical causal relation that is not a rung-3 causal relation, and in Example 9.2.14 we saw a rung-3 causal relation that is not a rung-2 causal relation.

We have formalized the main principle of how we can learn about causal relations in the world: by actively changing some part of the world (choosing the intervention values independently) and observing the response of other parts of the world. The independence assumption is key to distinguish mere correlation from causation.[55]

9.3. Direct Causal Relations

Another important notion is that of direct causation. Essentially, "a causes b directly" means that a causes b even when performing a hard intervention on all other endogenous variables. Since we distinguished different notions of "a causes b", we also end up distinguishing analogous notions of "a causes b directly".

One should keep in mind that the notion of direct causation is always relative to some set of variables.[56] In particular, this property is not necessarily preserved under marginalization.

We start again with the graphical notion. First, we need a graphical analogue of the notion of "submodel".

Definition 9.3.1. *Let $G = (J, V, E, L)$ be a CDMG with input nodes J and output nodes V. For $A \subseteq V$, we define the* ***submodel*** *of G on A as $G^{[A]} := G^{\mathrm{do}(V \setminus A)} = (J \cup (V \setminus A), A, \{v \to a \in E \,|\, v \in V, a \in A\}, \{a \leftrightarrow a' \in L \,|\, a, a' \in A\})$.*

We now define the graphical notion:

[55] As a less mathematical and more philosophical footnote: it is interesting to speculate about how this relates to the notion of a free will. If an agent is not convinced that it chose the intervention values independently of other past aspects of the world, it cannot validly perform this causal reasoning step. An agent without a free will to choose these values could therefore never conclude that its actions have a causal effect on the world, as it could also just be a puppet steered by higher powers, and any dependence it observes between its actions and aspects of the world could also be ascribed to confounding. So perhaps that is why evolution equipped us with the impression that we have a free will.

[56] This is implicitly communicated by the qualification "according to M" or "according to $G(M)$", but is easy to forget in practice when modeling some system causally, especially when one has not decided yet on the (complete) set of endogenous variables that need to be considered.

Definition 9.3.2. *Let G be a CDMG with input nodes J and output nodes V. Let $a \in J \cup V$ and $b \in V$. We say that a* ***is a direct cause of*** *b* ***(w.r.t.*** *$V \cup J$****) according to*** *G if a causes b according to $G^{[\{a,b\} \cap V]}$.*

Remark 9.3.3. *Note that this holds if and only if $a \in \mathrm{Pa}^G(b)$. This provides the causal interpretation of directed edges in a causal graph.*

Similarly, we define the rung-2 notion:

Definition 9.3.4. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let $a \in J \cup V$ and $b \in V$. We say that a* ***is a rung-2 direct cause of*** *b* ***(w.r.t.*** *$J \cup V$****) according to*** *M if and only if a is a rung-2 cause of b according to $M^{[\{a,b\} \cap V]}$.*

Remark 9.3.5. *Remember that a is a rung-2 cause of b according to M means:*

$$P_M(X_b \mid \mathrm{do}(X_{J \cup \{a\}})) \neq P_M(X_b \mid \mathrm{do}(X_{J \setminus \{a\}})).$$

Spelling this out, a being a rung-2 direct cause of b according to M means that:

$$P_M(X_b \mid \mathrm{do}(X_{(J \cup V) \setminus \{a,b\}}), \mathrm{do}(X_a)) \neq P_M(X_b \mid \mathrm{do}(X_{(J \cup V) \setminus \{a,b\}})).$$

In words: changing a influences the distribution of b even when intervening to hold all other endogenous variables fixed.

Remark 9.3.6. *Rung-2 direct causal relations are invariant under interventional equivalence of iSCMs. More precisely: a is a rung-2 direct cause of b (w.r.t. $J \cup V$) according to M implies that a is a rung-2 direct cause of b (w.r.t. $J \cup V$) according to $\tilde{M}$ if M is interventionally equivalent to $\tilde{M}$ (w.r.t. V).*

Similarly, we define the rung-3 notion:

Definition 9.3.7. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let $a \in J \cup V$ and $b \in V$. We say that a* ***is a rung-3 direct cause of*** *b* ***(w.r.t.*** *$J \cup V$****) according to*** *M if and only if a is a rung-3 cause of b according to $M^{[\{a,b\} \cap V]}$.*

Remark 9.3.8. *Remember that a is a rung-3 cause of b according to M means:*

$$\begin{aligned}&\exists x_{J \setminus \{I_a\}} \in \mathcal{X}_{J \setminus \{I_a\}}, \exists x_{I_a}, x_{I_{a'}} \in \mathcal{X}_{I_a} :\\ &P_{(M_{\mathrm{do}(X_{J \setminus \{I_a\}} = x_{J \setminus \{I_a\}}), \mathrm{do}(I_a)})^{\mathrm{twin}}}(X_b = X_{b'} \mid \mathrm{do}(X_{I_a} = x_{I_a}), \mathrm{do}(X_{I_{a'}} = x_{I_{a'}})) \neq 1.\end{aligned}$$

Spelling this out, a being a rung-3 direct cause of b according to M means that:

$$\begin{aligned}&\exists x_{(J \setminus \{I_a\}) \cup (V \setminus \{a,b\})} \in \mathcal{X}_{(J \setminus \{I_a\}) \cup (V \setminus \{a,b\})}, \exists x_{I_a}, x_{I_{a'}} \in \mathcal{X}_{I_a} :\\ &P_{(M_{\mathrm{do}(X_{(J \setminus \{I_a\}) \cup (V \setminus \{a,b\})} = x_{(J \setminus \{I_a\}) \cup (V \setminus \{a,b\})}), \mathrm{do}(I_a)})^{\mathrm{twin}}}(X_b = X_{b'} \mid \mathrm{do}(X_{I_a} = x_{I_a}), \mathrm{do}(X_{I_{a'}} = x_{I_{a'}})) \neq 1.\end{aligned}$$

Alternatively, we can spell this out using potential outcomes:

$$\begin{aligned}&\exists x_{(J \setminus \{I_a\}) \cup (V \setminus \{a,b\})} \in \mathcal{X}_{(J \setminus \{I_a\}) \cup (V \setminus \{a,b\})}, \exists x_{I_a}, x'_{I_a} \in \mathcal{X}_{I_a} :\\ &P_M\big(g_b(x_{I_a}, x_{(J \setminus \{I_a\}) \cup (V \setminus \{a,b\})}, X_W) = g_b(x'_{I_a}, x_{(J \setminus \{I_a\}) \cup (V \setminus \{a,b\})}, X_W)\big) \neq 1.\end{aligned}$$

where g is a solution function of $M^{[\{a,b\} \cap V]}_{\mathrm{do}(I_a)}$.

Remark 9.3.9. *Rung-3 direct causal relations are invariant under counterfactual equivalence of iSCMs. More precisely: a is a rung-3 direct cause of b (w.r.t. $J \cup V$) according to M implies that a is a rung-3 direct cause of b (w.r.t. $J \cup V$) according to $\tilde{M}$ if M is counterfactually equivalent to $\tilde{M}$ (w.r.t. V).*

We get a similar hierarchy of these notions:

Corollary 9.3.10. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM with causal graph $G(M)$. Let $a \in V \cup J$ and $b \in V$. Consider the following statements:*

(i) a is not a direct cause of b (w.r.t. $V \cup J$) according to $G(M)$;

(ii) a is not a rung-3 direct cause of b (w.r.t. $V \cup J$) according to M;

(iii) a is not a rung-2 direct cause of b (w.r.t. $V \cup J$) according to M.

Then (i) $\implies$ (ii) $\implies$ (iii). If $M^{[\{a,b\} \cap V]}_{\mathrm{do}(I_a)}$ is σ-faithful, also (iii) $\implies$ (i).

Proof. Apply Theorem 9.2.19 to $M^{[\{a,b\} \cap V]}$ and use that $G(M^{[\{a,b\} \cap V]}) = (G(M))^{[\{a,b\} \cap V]}$ (which follows from Proposition 6.9.6). □

Identifying a direct causal relation may not be very practical, as it seems to require intervening on *all* endogenous and exogenous input variables (except b) simultaneously. So, empirically, "a is a direct cause of b w.r.t. $J \cup V$" is a strong statement (the more variables are contained in $J \cup V$, the stronger it becomes).

9.4. Common Causes

Another important notion is that of "having a common cause". Essentially, "c is a common cause of a, b" means that c causes b even when performing a hard intervention on a, and c causes a even when performing a hard intervention on b. Equivalently, it means that c is a direct cause of both a and b with respect to $\{a, b, c\}$. Since we distinguished different notions of "a causes b", we also end up distinguishing analogous notions of "a and b have common cause c".

We start again with the graphical notion.

Definition 9.4.1. *Let G be a CDMG with input nodes J and output nodes V. Let $a, b \in V$ and $c \in V \cup J$ such that a, b, c are distinct. We say that c **is a common cause of** a **and** b **according to** G if c causes a according to $G_{\mathrm{do}(b)}$ and c causes b according to $G_{\mathrm{do}(a)}$.*

Remark 9.4.2. *Equivalent formulations are:*

- *if there exists a bifurcation with source c in G between a and b (use Proposition 3.2.4).*
- *if $a \leftarrow c$ and $c \rightarrow b$ are both present in $G^{\setminus (V \setminus \{a,b,c\})}$ (use Remark 3.2.19).*

Note that we do not consider exogenous random variables as candidate common causes. The reason is that we do not necessarily want to give these variables a causal interpretation (remember that we also did not formally define the effect of a hard intervention targeting such variables), in which case it would be misleading to refer to such a variable as a "common cause".

Similarly, we define the rung-2 notion:

Definition 9.4.3. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let $a, b \in V$ and $c \in V \cup J$ such that a, b, c are distinct. We say that c* ***is a rung-2 common cause of*** *a, b* ***according to*** *M if and only if c is a rung-2 cause of a according to $M_{\mathrm{do}(b)}$ and c is a rung-2 cause of b according to $M_{\mathrm{do}(a)}$.*

Remark 9.4.4. *Remember that c is a rung-2 cause of b according to M means:*

$$P_M(X_b \mid \mathrm{do}(X_{J\cup\{c\}})) \neq P_M(X_b \mid \mathrm{do}(X_{J\setminus\{c\}})).$$

Spelling this out, c being a rung-2 common cause of a, b according to M means that:

$$P_M(X_a \mid \mathrm{do}(X_{J\cup\{c\}}), \mathrm{do}(X_b)) \neq P_M(X_a \mid \mathrm{do}(X_{J\setminus\{c\}}), \mathrm{do}(X_b)) \text{ and}$$
$$P_M(X_b \mid \mathrm{do}(X_{J\cup\{c\}}), \mathrm{do}(X_a)) \neq P_M(X_b \mid \mathrm{do}(X_{J\setminus\{c\}}), \mathrm{do}(X_a)).$$

Remark 9.4.5. *The rung-2 notion of having a common cause is invariant under interventional equivalence of iSCMs. More precisely: c is a rung-2 common cause of a, b according to M implies that c is a rung-2 common cause of a, b according to $\tilde{M}$ if M is interventionally equivalent to $\tilde{M}$ w.r.t. $\{a, b, c\} \cap V$.*

Similarly, we define the rung-3 notion:

Definition 9.4.6. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let $a, b \in V$ and $c \in V \cup J$ such that a, b, c are distinct. We say that c* ***is a rung-3 common cause of*** *a, b* ***according to*** *M if and only if c is a rung-3 cause of a according to $M_{\mathrm{do}(b)}$ and c is a rung-3 cause of b according to $M_{\mathrm{do}(a)}$.*

Remark 9.4.7. *Remember that a is a rung-3 cause of b according to M means:*

$$\exists x_{J\setminus\{I_a\}} \in \mathcal{X}_{J\setminus\{I_a\}}, \exists x_{I_a}, x_{I_{a'}} \in \mathcal{X}_a :$$
$$P_{(M_{\mathrm{do}(X_{J\setminus\{I_a\}}=x_{J\setminus\{I_a\}}),\mathrm{do}(I_a)})^{\mathrm{twin}}}(X_b = X_{b'} \mid \mathrm{do}(X_{I_a} = x_{I_a}), \mathrm{do}(X_{I_{a'}} = x_{I_{a'}})) \neq 1.$$

Spelling this out, c being a rung-3 common cause of a, b according to M means that:

$$\exists x_{J\setminus\{I_c\}} \in \mathcal{X}_{J\setminus\{I_c\}}, \exists x_b \in \mathcal{X}_b, \exists x_{I_c}, x_{I_{c'}} \in \mathcal{X}_c :$$
$$P_{(M_{\mathrm{do}(X_{J\setminus\{I_c\}}=x_{J\setminus\{I_c\}}, X_b=x_b),\mathrm{do}(I_c)})^{\mathrm{twin}}}(X_a = X_{a'} \mid \mathrm{do}(X_{I_c} = x_{I_c}), \mathrm{do}(X_{I_{c'}} = x_{I_{c'}})) \neq 1.$$

and

$$\exists x_{J\setminus\{I_c\}} \in \mathcal{X}_{J\setminus\{I_c\}}, \exists x_a \in \mathcal{X}_a, \exists x_{I_c}, x_{I_{c'}} \in \mathcal{X}_c :$$
$$P_{(M_{\mathrm{do}(X_{J\setminus\{I_c\}}=x_{J\setminus\{I_c\}}, X_a=x_a),\mathrm{do}(I_c)})^{\mathrm{twin}}}(X_b = X_{b'} \mid \mathrm{do}(X_{I_c} = x_{I_c}), \mathrm{do}(X_{I_{c'}} = x_{I_{c'}})) \neq 1.$$

Alternatively, with potential outcomes this means:

$$\exists x_{J\setminus\{I_c\}} \in \mathcal{X}_{J\setminus\{I_c\}}, \exists x_b \in \mathcal{X}_b, \exists x_{I_c}, x'_{I_c} \in \mathcal{X}_c :$$
$$P_M(g_a(x_{I_c}, x_{J\setminus\{I_c\}}, x_b, X_W) = g_a(x'_{I_c}, x_{J\setminus\{I_c\}}, x_b, X_W)) \neq 1$$

where g is a solution function of $M_{\mathrm{do}(b),\mathrm{do}(I_c)}$, and

$$\exists x_{J\setminus\{I_c\}} \in \mathcal{X}_{J\setminus\{I_c\}}, \exists x_a \in \mathcal{X}_a, \exists x_{I_c}, x'_{I_c} \in \mathcal{X}_c :$$
$$P_M(g_b(x_{I_c}, x_{J\setminus\{I_c\}}, x_a, X_W) = g_b(x'_{I_c}, x_{J\setminus\{I_c\}}, x_a, X_W)) \neq 1.$$

where g is a solution function of $M_{\mathrm{do}(a),\mathrm{do}(I_c)}$.

Remark 9.4.8. *The rung-3 notion of having a common cause is invariant under counterfactual equivalence of iSCMs. More precisely: c is a rung-3 common cause of a, b according to M implies that c is a rung-3 common cause of a, b according to $\tilde{M}$ if M is counterfactually equivalent to $\tilde{M}$ w.r.t. $\{a, b, c\} \cap V$.*

We get a similar hierarchy of these notions:

Corollary 9.4.9. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM with causal graph $G(M)$. Let $a, b \in V$ and $c \in V \cup J$ such that a, b, c are distinct. Consider the following statements:*

(i) c is not a common cause of a, b according to $G(M)$;

(ii) c is not a rung-3 common cause of a, b according to M;

(iii) c is not a rung-2 common cause of a, b according to M.

Then (i) $\implies$ (ii) $\implies$ (iii). If $M_{\mathrm{do}(b),\mathrm{do}(I_c)}$ and $M_{\mathrm{do}(a),\mathrm{do}(I_c)}$ are both σ-faithful, then (iii) $\implies$ (i).

Proof. Apply Theorem 9.2.19 to $M_{\mathrm{do}(a)}$ and $M_{\mathrm{do}(b)}$, and use Proposition 6.9.6. □

We could go on to define "direct common causes" (in analogy to how we defined "direct causes" in terms of "causes") but we will not do so here, as this concept does not seem to be of much importance.

9.5. Confounding

The important notion of "confounding" or "spurious dependence" is somewhat tricky to define, especially if we do not exclude the possibility of cycles. Roughly speaking, we say that two endogenous variables are confounded if their joint distribution is not entirely explained by their mutual causal effects.

9.5.1. Confounding (graphical notion)

Definition 9.5.1. *Let G be a CDMG with input nodes J and output nodes V. Let $a, b \in V$ be distinct output nodes. If there exists a bifurcation between a and b in G, either without source or with source $c \in V$, then we say that a* ***and*** *b* ***are confounded according to*** *G.*[57] *Otherwise, we say that a* ***and*** *b* ***are unconfounded according to*** *G.*

Remark 9.5.2. *Marginalization of a graph preserves confoundedness (see also Remark 3.2.19). In particular, a and b are confounded according to G if and only if $a \leftrightarrow b$ is present in $G^{\setminus(V\setminus\{a,b\})}$.*

We can find an (almost) equivalent formulation of unconfoundedness as a separation statement by using the twinning operation.[58]

Proposition 9.5.3. *Let G^+ be a CDMG with input nodes J and output nodes $V \cup W$ such that the nodes in W have no parents in G^+. Let $a, b \in V$ be distinct output nodes in V. If a and b are unconfounded according to $G := (G^+)^{\setminus W}$, then*

$$a \underset{(G^+)^{\mathrm{twin}(J\cup V)}_{\mathrm{do}(a',b)}}{\overset{\sigma}{\perp}} b' \mid J, J', a', b.$$

The converse holds if each node in V has at least one parent in W in G^+.

Proof. $\Longrightarrow$: Suppose there exists a σ-open walk in the graph $(G^+)^{\mathrm{twin}(J\cup V)}_{\mathrm{do}(a',b)}$ between a and $b' \cup J \cup J' \cup \{a', b\}$. Then there exists a σ-open walk in the graph $(G^+)^{\mathrm{twin}(J\cup V)}_{\mathrm{do}(a',b)}$ between a and $b' \cup J \cup J' \cup \{a', b\}$. There must be such a walk of minimal length. In particular, it contains a and b' both at most once. It cannot contain nodes from $J \cup J' \cup \{a', b\}$, as such a node would either be an end node of the walk (σ-blocking it) or a non-collider node pointing only to nodes in another strongly connected component (σ-blocking the walk). Hence the walk must be between a and b'. It cannot contain any collider. It cannot be a directed walk because it has to pass through a node in W corresponding to an exogenous random variable in M, but those nodes have no incoming edges. Therefore it must be a walk of the form $a \leftarrow \cdots \leftarrow w \rightarrow \cdots \rightarrow b'$ with $w \in W$, because only nodes in W can be ancestors of endogenous nodes in different "worlds". Note that the subwalk $a \leftarrow \cdots \leftarrow w$ must consist of nodes in $V \cup \{w\}$ and cannot contain b. Also, the subwalk $w \rightarrow \cdots \rightarrow b'$ must consist of nodes in $V' \cup \{w\}$ and cannot contain a'. By "removing the primes" from the nodes in this subwalk, we obtain a walk of the form $a \leftarrow \cdots \leftarrow w \rightarrow \cdots \rightarrow b$ in $G^+(M)$, which is seen to be a bifurcation with

[57] The reason we exclude bifurcations with as source an exogenous input node is that these will not lead to "spurious dependences" in the "strata" of the Markov kernel.

[58] The only problem may arise is that a and b are confounded because of a common cause c that has no exogenous random parents. Strictly speaking, we would not want to consider this as confounding (because if c is deterministic, it will induce no spurious dependence). This is not captured adequately by our graphical criterion simply because the causal graph where the exogenous random variables have been marginalized out does not encode this piece of information.

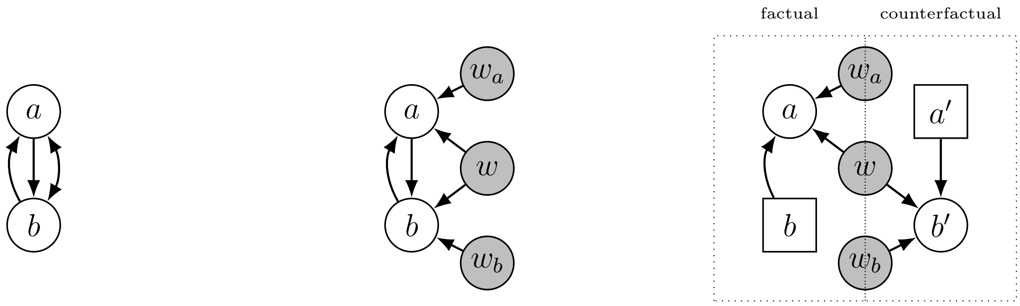


Figure 25: Left: Marginal graph $(G^+)^{\setminus W}$. Center: Graph G^+. Right: Twinned intervened $(G^+)^{\mathrm{twin}(J\cup V)}_{\mathrm{do}(a',b)}$. Proposition 9.5.3 states that if a and b are unconfounded according to $(G^+)^{\setminus W}$, then there will be no σ-open path between a and b' given a', b in the twinned intervened graph.

source $w \in W$ in G^+. But then there exists a bifurcation in $(G^+)^{\setminus W}$ between a and b without source, contradicting the assumptions.

$\Longleftarrow$: if a and b are confounded according to G then there exists a bifurcation between a and b in G, either without source, or with source $c \in V$. In the former case, there exists a bifurcation between a and b in G^+ with source $w \in W$. In the latter case, if each node in V has at least one parent in W in G^+, then there also exists a bifurcation between a and b in G^+ with source $w \in W$. The bifurcation is of the form $a \leftarrow \cdots \leftarrow w \rightarrow \cdots \rightarrow b$. It only consists of nodes in V, except for $w \in W$. We can add primes to all nodes on the subwalk from w to b (except on w) to obtain a walk of the form $a \leftarrow \cdots \leftarrow w \rightarrow \cdots \rightarrow b'$. Note that the subwalk $a \leftarrow \cdots \leftarrow w$ must consist of nodes in $V \cup \{w\}$ and cannot contain b. Also, the subwalk $w \rightarrow \cdots \rightarrow b'$ must consist of nodes in $V' \cup \{w\}$ and cannot contain a'. This is a σ-open walk between a and b' given $J \cup J' \cup \{a', b\}$. $\square$

9.5.2. Confounding (counterfactual/potential outcome notion)

Definition 9.5.4. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let $a, b \in V$ with $a \neq b$. Let $g^{\mathrm{do}(a)}$ be a solution function of $M_{\mathrm{do}(a)}$, and $g^{\mathrm{do}(b)}$ a solution function of $M_{\mathrm{do}(b)}$. If for all values $x_J \in \mathcal{X}_J, x_{J'} \in \mathcal{X}_J$, $x_{a'} \in \mathcal{X}_a$, $x_b \in \mathcal{X}_b$:*

$$g_a^{\mathrm{do}(b)}(x_b, x_J, X_W) \perp\!\!\!\perp g_b^{\mathrm{do}(a)}(x_{a'}, x_{J'}, X_W)$$

with $X_W \sim P$, then we say that a ***and*** *b* ***are rung-3 unconfounded according to*** *M.*

Remark 9.5.5. *This is equivalent to:*

$$X_a \underset{P_{M^{\mathrm{twin}}_{\mathrm{do}(a',b)}}}{\perp\!\!\!\perp} X_{b'} \,|\, X_J, X_{J'}, X_{a'}, X_b.$$

Indeed, for all $x_J \in \mathcal{X}_J, x_{J'} \in \mathcal{X}_J$, $x_{a'} \in \mathcal{X}_a$, $x_b \in \mathcal{X}_b$:

$$\begin{aligned}&\big(g_a^{\mathrm{do}(b)}(x_b, x_J, X_W), g_b^{\mathrm{do}(a)}(x_{a'}, x_{J'}, X_W)\big)\\ &\quad\sim P_{M^{\mathrm{twin}}_{\mathrm{do}(a',b)}}(X_a, X_{b'} \,|\, \mathrm{do}(X_J = x_J, X_{J'} = x_{J'}, X_{a'} = x_{a'}, X_b = x_b)).\end{aligned}$$

Therefore, the choice of the solution functions and the random variable X_W do not matter in Definition 9.5.4.

Remark 9.5.6. *Rung-3 unconfoundedness is invariant under counterfactual equivalence of iSCMs. More precisely: a and b are rung-3 unconfounded according to M implies that a and b are rung-3 unconfounded according to $\tilde{M}$ if M is counterfactually equivalent to $\tilde{M}$ w.r.t. $\{a, b\}$.*

However, the notion of rung-3 unconfoundedness is *not* invariant under interventional equivalence.

Example 9.5.7. *Consider again the SCM M in Example 9.2.9. It has endogenous variables $X_a, X_b \in \{-1, +1\}$, exogenous random variable $X_w \in \{-1, 1\}$, and structural equations:*

$$\begin{aligned}X_a &= X_w,\\ X_b &= X_a X_w,\end{aligned}$$

where $X_w \sim \mathcal{U}(\{-1, 1\})$.

According to M, a and b are rung-3 confounded (indeed, the potential outcomes have perfect dependence). M is interventionally equivalent to SCM $\tilde{M}$ with endogenous variables $X_a, X_b \in \{-1, +1\}$, exogenous random variables $X_u, X_w \in \{-1, 1\}$, and structural equations:

$$\begin{aligned}X_a &= X_u,\\ X_b &= X_a X_w\end{aligned}$$

where $X_u, X_w \sim \mathcal{U}(\{-1, 1\})$ are independent. According to $\tilde{M}$, a and b are rung-3 unconfounded.

Proposition 9.5.8. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM with causal graph $G(M)$. Let $a, b \in V$ with $a \neq b$. If a and b are unconfounded according to $G(M)$ then a and b are rung-3 unconfounded according to M.*

Proof. If a and b are unconfounded according to $G(M)$, then

$$a \underset{(G^+(M))^{\mathrm{twin}(V,J)}_{\mathrm{do}(a',b)}}{\overset{\sigma}{\perp}} b' \,|\, J, J', a', b$$

by Proposition 9.5.3. Since $(G^+(M))^{\mathrm{twin}(V,J)}_{\mathrm{do}(a',b)} = G^+(M^{\mathrm{twin}})_{\mathrm{do}(a',b)} = G^+(M^{\mathrm{twin}}_{\mathrm{do}(a',b)})$, we get

$$a \underset{G(M^{\mathrm{twin}}_{\mathrm{do}(a',b)})}{\overset{\sigma}{\perp}} b' \,|\, J, J', a', b.$$

By the global Markov property applied to $M^{\text{twin}}_{\text{do}(a',b)}$, we obtain that

$$X_a \underset{P_{M^{\text{twin}}_{\text{do}(a',b)}}}{\perp\!\!\!\perp} X_{b'} \mid X_J, X_{J'}, X_{a'}, X_b.$$

The claim follows from Remark 9.5.5. □

9.5.3. Confounding (acyclic potential outcome notion)

In the potential outcome literature, one also encounters other definitions of unconfoundedness that are formulated in terms of potential outcomes. These can only be applied if a and b are not part of a cycle, and hence this notion is less general than Definition 9.5.4. We will show how these notions can be seen as a special case of Definition 9.5.4, under the assumption that the potential outcomes are induced by a simple iSCM.

Proposition 9.5.9. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let $a, b \in V$ with $a \neq b$. Suppose that b is not a rung-3 cause of a according to M. Then a and b are rung-3 unconfounded according to M if and only if*

$$\forall x_J, x'_J \in \mathcal{X}_J \; \forall x'_a \in \mathcal{X}_a : g_a(x_J, X_W) \perp\!\!\!\perp g_b^{\text{do}(a)}(x'_a, x'_J, X_W) \tag{52}$$

for $X_W \sim P(X_W)$ (where $g^{\text{do}(a)}$ denotes a solution function of $M_{\text{do}(a)}$, and g a solution function of M).

Proof. Let $X_W \sim P(X_W)$. If b is not a rung-3 cause of a according to M, then for all $x_J \in \mathcal{X}_J$, for all $x_b \in \mathcal{X}_b$, for $P(X_W)$-almost all $x_W \in \mathcal{X}_W$:

$$g_a^{\text{do}(b)}(x_b, x_J, x_W) = g_a(x_J, x_W).$$

Hence,

$$\forall x_J, x'_J \in \mathcal{X}_J \; \forall x'_a \in \mathcal{X}_a : g_a(x_J, X_W) \perp\!\!\!\perp g_b^{\text{do}(a)}(x'_a, x'_J, X_W)$$

if and only if

$$\forall x_J, x'_J \in \mathcal{X}_J \; \forall x'_a \in \mathcal{X}_a \; \forall x_b \in \mathcal{X}_b : g_a^{\text{do}(b)}(x_b, x_J, X_W) \perp\!\!\!\perp g_b^{\text{do}(a)}(x'_a, x'_J, X_W).$$

□

Remark 9.5.10. *(52) is equivalent to:*

$$X_a \underset{P_{M^{\text{twin}}_{\text{do}(a')}}}{\perp\!\!\!\perp} X_{b'} \mid X_J, X'_J, X'_a.$$

Indeed, for all $x_J \in \mathcal{X}_J, x'_J \in \mathcal{X}_J$, $x'_a \in \mathcal{X}_a$:

$$\begin{aligned}&\left(g_a(x_J, X_W), g_b^{\text{do}(a)}(x'_a, x'_J, X_W)\right)\\ &\quad \sim P_{M^{\text{twin}}_{\text{do}(a')}}(X_a, X_{b'} \mid \text{do}(X_J = x_J, X_{J'} = x_{J'}, X_{a'} = x'_a)).\end{aligned}$$

We now relate this to the assumptions regarding unconfoundedness that are usually made in the potential outcome literature. We consider the case $J = \emptyset$. Let M be an SCM satisfying the assumptions of Proposition 9.5.9. Let X_W be a random variable with distribution $X_W \sim P$. Let $g^{\mathrm{do}(a)}$ be the solution function of $M_{\mathrm{do}(a)}$, and g the solution function of M. We can then define potential outcomes X_a for M and $X_b^{\mathrm{do}(x_a)}$ for $M_{\mathrm{do}(a)}$ (the latter with input x_a):

$$
\begin{aligned}
X_a &:= g_a(X_W) \\
X_b^{\mathrm{do}(x_a)} &:= g_b^{\mathrm{do}(a)}(x_a, X_W)
\end{aligned}
$$

for all $x_a \in \mathcal{X}_a$.[59] Equation (52) can then also be written in terms of these potential outcomes as

$$\forall x_a \in \mathcal{X}_a : \qquad X_a \perp\!\!\!\perp X_b^{\mathrm{do}(x_a)}.$$

In the more common notation, for binary treatment $T := X_a \in \{0,1\}$ and corresponding potential outcomes $Y(t) := X_b^{\mathrm{do}(x_a)}$ with $x_a = t$, this reads as:

$$\forall t \in \{0,1\} : \qquad T \perp\!\!\!\perp Y(t).$$

This property (of the treatment variable and the corresponding potential outcome variables) is referred to as "exchangeability" or "ignorability" in the potential outcome framework, and expresses within that framework the assumption of "no confounding" when the task is to estimate the causal effect of a on b. One should keep in mind the assumption that b is not a rung-3 cause of a. In this setting, this can be written as:

$$\forall y \in \mathcal{Y} : \qquad T(y) = T \text{ a.s.},$$

where we wrote $\mathcal{Y} := \mathcal{X}_b$ for the possible values of the outcome, and denoted "potential treatments" $T(y) := X_a^{\mathrm{do}(X_b = x_b)} := g_a^{\mathrm{do}(b)}(x_b, X_W)$ with $x_b = y$.

9.5.4. Confounding (acyclic interventional notion)

Another notion of confounding is defined at the level of (interventional) distributions. If we have a "treatment" and an "outcome variable" (and assume that outcome does not cause treatment), then we consider them to have "no confounding bias" if the distribution of the outcome does not depend on whether we just observed the treatment, or if we intervened to impose the treatment. We provide here the definition only in case the treatment and outcome variable are not part of a causal cycle, leaving the question of how to properly define this notion in full generality for simple iSCMs as an open research question.

Definition 9.5.11. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let $a \neq b \in V$. Assume b is not a rung-2 cause of a according to M. Then we say that a* ***and*** *b* ***have no confounding bias according to*** *M if:*

$$P_M(X_b \mid \mathrm{do}(X_a), \mathrm{do}(X_J)) = P_M(X_b \mid X_a, \mathrm{do}(X_J)) \qquad P_M(X_a \mid \mathrm{do}(X_J))\text{-}a.s.. \tag{53}$$

[59] Note that we use the same exogenous random variable X_W to "couple" these potential outcomes.

Colloquially, confounding bias is also often referred to as "spurious dependence". Also, we could say that there is no confounding bias if and only if "causation equals correlation". If we train a prediction model that estimates $\mathbb{E}(X_b \,|\, X_a)$, and we know that X_b does not cause X_a and X_a and X_b have no confounding bias, then we can use the model for making causal predictions, that is, for estimating $\mathbb{E}(X_b \,|\, \mathrm{do}(X_a))$. The assumption that two variables have no confounding bias is often (perhaps too often) made in practice.[60]

Remark 9.5.12. *We can write (53) equivalently as:*

$$X_b \underset{P_{M_{\mathrm{do}(I_a)}}}{\perp\!\!\!\perp} X_{I_a} \,|\, X_{\{a\}\cup J}.$$

Indeed, this conditional independence means that there exists a Markov kernel $\mathcal{X}_a \times \mathcal{X}_J \dashrightarrow \mathcal{X}_b$ that is a version of $P_M(X_b \,|\, X_a, \mathrm{do}(X_J))$ and $P_M(X_b \,|\, \mathrm{do}(X_a), \mathrm{do}(X_J))$ simultaneously (see proof of Proposition 5.1.8). This has to be $P_M(X_b \,|\, \mathrm{do}(X_a), \mathrm{do}(X_J))$.

Remark 9.5.13. *"Having no confounding bias" is invariant under interventional equivalence of iSCMs. More precisely: if a and b have no confounding bias according to M, then a and b have no confounding bias according to $\tilde{M}$ if M is interventionally equivalent to $\tilde{M}$ w.r.t. $\{a, b\}$.*

Proposition 9.5.14. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM. Let $a, b \in V$ be distinct endogenous variables such that b is not a rung-3 cause of a according to M. If a and b are rung-3 unconfounded according to M, then a and b have no confounding bias according to M.*

Proof. Let $X_W \sim P(X_W)$. Proposition 9.5.9 states:

$$\forall x_J, x'_J \in \mathcal{X}_J \forall x'_a \in \mathcal{X}_a : \qquad g_a(x_J, X_W) \perp\!\!\!\perp g_b^{\mathrm{do}(a)}(x'_a, x'_J, X_W).$$

Consistency (Proposition 7.5.1) of $M_{\{a,b\}}$ implies:

$$\forall x_J \in \mathcal{X}_J \forall x_W \in \mathcal{X}_W : g_b^{\mathrm{do}(a)}(g_a(x_J, x_W), x_J, x_W) = g_b(x_J, x_W).$$

Hence, for all $x_J \in \mathcal{X}_J$, for $P(g_a(x_J, X_W))$-almost every $x_a \in \mathcal{X}_a$:

$$\begin{aligned} P(g_b(x_J, X_W) \,|\, g_a(x_J, X_W) = x_a) &= P(g_b^{\mathrm{do}(a)}(g_a(x_J, X_W), x_J, X_W) \,|\, g_a(x_J, X_W) = x_a) \\ &= P(g_b^{\mathrm{do}(a)}(x_a, x_J, X_W) \,|\, g_a(x_J, X_W) = x_a) \\ &= P(g_b^{\mathrm{do}(a)}(x_a, x_J, X_W)). \end{aligned}$$

Hence, in the usual notation:

$$P_M(X_b \,|\, \mathrm{do}(X_J), \mathrm{do}(X_a)) = P_M(X_b \,|\, \mathrm{do}(X_J), X_a) \qquad P_M(X_a \,|\, \mathrm{do}(X_J))\text{-a.s.}$$

□

[60] If one doesn't rule out possible confounding bias, one can still obtain bounds on causal predictions, see Theorem 7.5.2.

This yields the following criterion to detect the existence of certain bifurcations in $G(M)$:

Corollary 9.5.15. *Let $M = (J, V, W, \mathcal{X}, P, f)$ be a simple iSCM with causal graph $G(M)$. Let $a, b \in V$ with $a \neq b$. Assume that b does not cause a according to $G(M)$. If a and b have confounding bias according to M, then a and b must be confounded according to $G(M)$.*

Proof. Combine Proposition 9.5.14, Theorem 9.2.19 and Proposition 9.5.8. □

This condition for identifying confoundedness in a causal graph is sufficient, but not necessary, as the next example shows.

Example 9.5.16. *Consider the (acyclic) SCM M with binary endogenous variables X_1, X_2, $X_3 \in \{-1, 1\}$ and a single exogenous random variable $E = (E_1, E_2, E_3) \in \{-1, 1\}^3$, structural equations:*

$$\begin{aligned} X_1 &= E_1 \\ X_2 &= E_2 \\ X_3 &= E_3, \end{aligned}$$

and exogenous distribution given by $P(E = e) = \frac{1}{4}\delta_{e_3 = e_1 e_2}$.

Consider now the (acyclic) SCM $\tilde{M}$ with binary endogenous variables $X_1, X_2, X_3 \in \{-1, 1\}$ and two exogenous random variables $E_1, E_2 \in \{-1, 1\}$, and structural equations:

$$\begin{aligned} X_1 &= E_1 \\ X_2 &= E_2 \\ X_3 &= E_1 E_2, \end{aligned}$$

where $E_1, E_2 \sim \mathcal{U}(\{-1, 1\})$ are independent.

The causal graphs of M and $\tilde{M}$ are depicted in Figure 26. M and $\tilde{M}$ are interventionally equivalent. Indeed, they are even counterfactually equivalent; this follows since the change-of-variables $(E_1, E_2, E_3) \mapsto (\tilde{E}_1, \tilde{E}_2) := (E_1, E_2)$ is an exogenous push-forward (that does not depend on exogenous input variables). The bidirected edge $X_1 \leftrightarrow X_2$ in $G(M)$ is not present in $G(\tilde{M})$ (see Figure 26). Analogously to $\tilde{M}$, one can define counterfactually equivalent iSCMs that have another combination of two out of the three bidirected edges of $G(M)$. Note that X_1 and X_2 have no confounding bias according to each of these iSCMs (as are X_1 and X_3, and X_2 and X_3). Still, within the counterfactual equivalence class, these three iSCMs with two bidirected edges each have a minimal causal graph. In other words, there exists no counterfactually equivalent iSCM with less than two bidirected edges and without any directed edges in its causal graph.

This example shows that it is possible that no pair of endogenous variables has confounding bias, although bidirected edges are still necessary to represent exogenous dependences between endogenous variables.[61]

[61] This suggests to extend the definition of confounding bias to a group of variables rather than just a pair of variables, but we will not do so here.

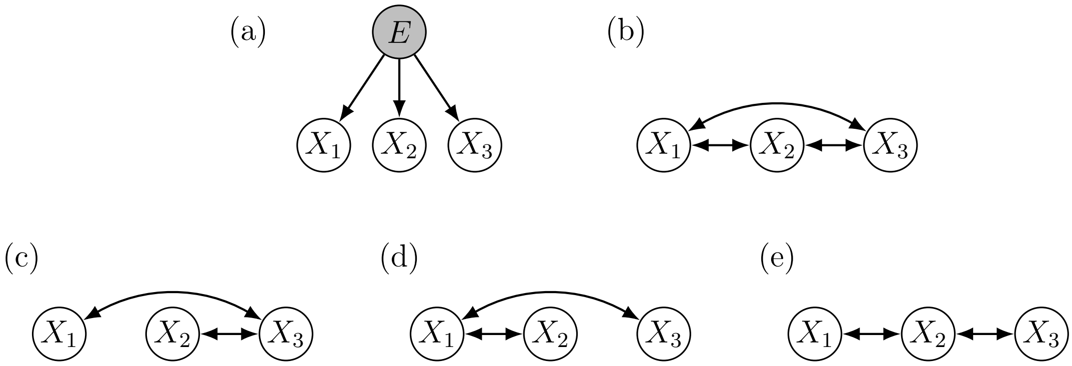


Figure 26: Example 9.5.16: (a) $G^{+}(M)$; (b): $G(M)$; (c): $G(\tilde{M})$; (d-e) causal graphs of two other iSCMs counterfactually equivalent to M.

Remark 9.5.17. *How to define "a and b have confounding bias" if a and b are part of a causal cycle is at present an open research problem.*

10. Causal Discovery & Estimation with iSCMs

So far, we always assumed that an iSCM was fully specified, and derived theory to draw conclusions from the given iSCM. For example, the do-calculus provides precise relationships between certain Markov kernels induced by the iSCM, which enables us to perform *causal reasoning*.

However, often we do not have sufficient information regarding the system that we are modeling to completely specify an iSCM. For example, we may only know what the observed variables are, but not what the graph of the iSCM is, let alone know the latent spaces, exogenous distribution and exact causal mechanisms. Can we still perform causal reasoning with such incompletely specified models? The answer turns out to be affirmative, if one is willing to make certain assumptions (that—unfortunately—are typically untestable, even when having access to the observable Markov kernel of the model).

In the rest of this chapter we will focus on the question of how to deduce partial knowledge about the iSCM from given Markov kernels. When this knowledge pertains only to the causal graph of the iSCM, this is often called *causal discovery*. When it pertains to parameters of the iSCM, for example, when estimating the causal effect of a variable on another, this is often referred to as *causal inference* (although "inference" is often interpreted much broader as drawing conclusions from data and prior beliefs).

In the next chapter, we will go one step further, and replace the deduction of iSCM properties from Markov kernels by the estimation of iSCM properties from data, i.e., we replace Markov kernels by finite samples. This will lead to statistical considerations. In particular, we will focus on estimating the conditional independences of an iSCM from data.

In this chapter, we will make use of the simple iSCM formalism. Similar (more restrictive) results can be obtained in the L-CBN formalism.

10.1. Randomized Controlled Trials

The notion of randomized controlled trials (also known as A/B-testing in engineering), is centuries old. It was already proposed by the Flemish physician Jan Baptista van Helmont [vH48] in 1648. As of today, it still provides the 'gold standard' for discovering causal relations and for the estimation of causal effects.

The experimental procedure is as follows. Consider two variables, "treatment" C and "outcome" X. In the simplest setting, one considers a binary treatment variable, where $C = 1$ corresponds to "treat with drug" and $C = 0$ corresponds to "treat with placebo" in a medical setting, or with "arm A" and "arm B" in an engineering setting. For example, the drug could be aspirin, and outcome could be the severity of headache perceived two hours later. Patients are split into two groups, the treatment and the control group, by means of a coin flip that assigns a value of C to every patient.[62] Patients are treated

[62]Usually this is done in a double-blind way, so that neither the patient nor the doctor knows which group a patient has been assigned to.

depending on the assigned value of C, i.e., patients in the treatment group are treated with the drug and patients in the control group are treated with a placebo. Some time after treatment, the outcome X is measured for each patient. This yields a data set $(C_n, X_n)_{n=1}^N$ with two measurements (C_n, X_n) for the n^{th} patient. If the distribution of outcome X significantly differs between the two groups, one concludes that treatment is a cause of outcome.

Let us formalize this in the causal modeling language of iSCMs. Apart from that treatment may have a causal effect on outcome, there are likely many other factors that influence outcome. Some have been measured, others not. For obvious practical reasons, we are not going to explicitly model *all* of them. Formally, we will assume that an accurate causal model of the situation is provided by some (unknown) simple iSCM with observed variables C and X, and possibly other latent variables. We will consider the outcome variable X as endogenous. But what type of variable should we consider the treatment variable C to be (which is not necessarily binary)? We have two possibilities: exogenous input, or endogenous. We will discuss both of these possibilities in sequence.

Let us start by considering the treatment variable C as an exogenous input variable. This choice encodes into the iSCM that outcome X does not cause treatment C. We are interested in answering two questions. The first is "Does treatment cause outcome according to $G(M)$?", where M is the underlying (unknown) iSCM. Since there are only two observed variables, this is equivalent to asking "Is $C \rightarrow X$ in $G(M)$?". The second question is "What is the causal effect of treatment on outcome?". We interpret this as asking for the Markov kernel $P_M(X \mid \mathrm{do}(C))$.

Proposition 10.1.1. *Let M be a simple iSCM with a single exogenous input variable C and a single endogenous variable X. A dependence*

$$X \underset{P_M(X|\mathrm{do}(C))}{\not\perp\!\!\!\perp} C \tag{54}$$

implies that C causes X according to $G(M)$.

Proof. This follows immediately from Theorem 9.2.19. □

The second question is, in this case, trivially answered.

The other option is to consider the treatment variable as *endogenous*. One situation in which this is more appropriate is so-called "imperfect compliance". If trial subjects do not all comply with prescribed treatment, for whatever reasons, then we can no longer identify the coin flip outcome with treatment, (even though coin flip outcome may still be an important cause of treatment).

In this modeling variant, we assume the existence of a simple iSCM M with endogenous variables C, X, both of which are observed (and possibly other latent variables). C here retains the meaning of treatment (but is no longer necessarily identifiable with the coin flip result). Under additional assumptions regarding the exogeneity of the treatment variable, we again obtain a similar statement as before.

Proposition 10.1.2. *Let $\tilde{M}$ be a simple SCM with two observed endogenous variables C, X, and no exogenous input variables. Under the following two assumptions:*

1. *X does not cause C according to $G(\tilde{M})$, and*

2. *C and X are unconfounded according to $G(\tilde{M})$,*

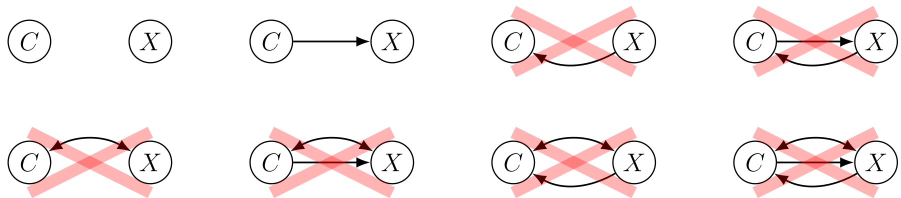
a dependence

$$X \underset{P_{\tilde{M}}(X,C)}{\not\perp\!\!\!\perp} C \tag{55}$$

implies that C causes X according to $G(\tilde{M})$, and the causal effect of C on X satisfies:

$$P_{\tilde{M}}\big(X \,|\, \mathrm{do}(C)\big) = P_{\tilde{M}}(X \,|\, C) \qquad P_{\tilde{M}}(C)\text{-a.s.}. \tag{56}$$

Proof. The first assumption is equivalent to $C \leftarrow X \notin G(\tilde{M})$, and the second assumption is equivalent to $C \leftrightarrow X \notin G(\tilde{M})$. Hence, out of the eight possible graphs $G(\tilde{M})$, only two satisfy the assumptions:



By the Markov property, if the edge $C \to X$ were absent in $G(\tilde{M})$, then $X \perp\!\!\!\perp_{P_{\tilde{M}}(X,C)} C$. In both cases, rule 2 of the causal do-calculus applied to $\tilde{G}$ yields the identity (56). □

Equation (54) is equivalent to the existence of values $c, c' \in \mathcal{X}_C$ such that

$$P_M(X \,|\, \mathrm{do}(C = c)) \neq P_M(X \,|\, \mathrm{do}(C = c')).$$

Equation (55) is equivalent to the existence of values

$$P_{\tilde{M}}(X \,|\, C = c) \neq P_{\tilde{M}}(X \,|\, C = c')$$

for every version of $P_{\tilde{M}}(X \,|\, C)$. These two statements are subtly different. We will see in the next chapter, that as long as C is discrete, they are actually not that different when testing these statements from a finite sample.

Apart from assuming that there exists a simple iSCM that provides an accurate model, in both cases, we made the following (implicit or explicit) causal assumptions regarding the treatment variable:

1. outcome X does not cause treatment C according to $G(M)$;

2. outcome X and treatment C are unconfounded according to $G(M)$.

As a first remark, we opted here for the strongest version of the notions "does not cause" and "are unconfounded"—the graphical ones. This might not be *necessary* for deriving the desired conclusions, but it surely is convenient.[63] The first assumption is commonly deemed justified if the outcome is an event that occurs later in time than the treatment event.[64] The second assumption is usually defended by appealing to randomization. Indeed, if treatment is decided solely by a coin flip, then it seems hard to imagine some factor that can influence the outcome of the coin flip and also influence the outcome X.[65] If in addition it is ensured that all subjects are included in the data set (to avoid introducing selection bias), then it is hard to conceive of any remaining source of confounding bias between outcome X and treatment C.[66]

We have shown (in two slightly different ways) that under these assumptions, if the distribution of the outcome X differs between the two groups of patients ("treatment group" with $C = 1$ vs. "control group" with $C = 0$), then treatment must be a cause of the outcome, at least in this population of patients.

If one formally considers treatment as an exogenous input variable, but then also assumes that its values are randomly assigned, then the differences between both modeling approaches are purely cosmetical. However, there is one advantage that the exogenous input approach has over the other approach: it does not restrict in any way the possible values of the treatment variable. This allows more freedom in the experimental design and sampling scheme design. For example, one can decide prior to conducting the RCT that the sampling scheme should end up with an equal number of patients in both groups. In case treatment is assigned by flipping a coin for each patient, it is rather unlikely that we will end up with exactly the same number of patients in both groups.

10.2. Estimating average treatment effects

Another way to formalize the randomized controlled setting is by using potential outcomes. For a binary treatment variable, one introduces two random variables per patient: $X_n^{\mathrm{do}(c_n=1)}$ and $X_n^{\mathrm{do}(c_n=0)}$, corresponding to the potential outcomes for the n'th patient if we treat the patient, or not, respectively. Given the actual treatment C_n, we then define the actual outcome as $X_n := X_n^{\mathrm{do}(c_n=C_n)}$. In practice, we only observe the actual outcome for each patient, while the other potential outcome for each patient remains latent.

An estimand that one is typically interested in and that captures the average strength of the causal effect of the treatment on the outcome is the so-called *average treatment*

[63] Some researchers prefer potential outcomes because these do not require them to make such graphical assumptions, or even to talk about a graph at all. In our opinion, causal graphs are very convenient to communicate causal modeling assumptions, and we are willing to pay the price for slightly strengthening the untestable assumptions from rung-3 to the graphical rung.

[64] The only exception seems to be certain scenarios involving time travel in science fiction novels/movies.

[65] Perhaps an almighty God could be such a factor, though this possibility is often met with skepticism in scientific circles.

[66] An example in which selection bias would be present is if patients in the placebo group that become too ill drop out of the study (for example, for ethical reasons) and are not included in the data.

effect (ATE)

$$\tau := \mathbb{E}(X_n^{\mathrm{do}(c_n=1)} - X_n^{\mathrm{do}(c_n=0)}).$$

To estimate the ATE, one assumes that C_n is randomized. This motivates the assumption that treatment and outcome are unconfounded:

$$X_n^{\mathrm{do}(c_n=0)} \perp\!\!\!\perp C_n, \qquad X_n^{\mathrm{do}(c_n=1)} \perp\!\!\!\perp C_n$$

(which follows from Proposition 9.5.9 if one assumes an underlying iSCM). One can then show that the difference-in-means estimator

$$\hat{\tau} := \frac{1}{|\{n : C_n = 1\}|} \sum_{\substack{n=1 \\ C_n=1}}^{N} X_n - \frac{1}{|\{n : C_n = 0\}|} \sum_{\substack{n=1 \\ C_n=0}}^{N} X_n$$

is an unbiased, consistent estimator of the ATE τ. Curiously enough, while we can speak of the difference $X_n^{\mathrm{do}(c_n=1)} - X_n^{\mathrm{do}(c_n=0)}$ as the individual treatment effect, this is in many settings an unobservable quantity; however, the *average* treatment effect can be estimated from observed data. In the iSCM setting, we can think of the potential outcomes as counterfactuals in a twin iSCM. However, when assuming an underlying iSCM, there is no need to go to the counterfactual level, as one can simply define the ATE as

$$\tau := \mathbb{E}_M(X \mid \mathrm{do}(C = 1)) - \mathbb{E}_M(X \mid \mathrm{do}(C = 0)).$$

In the presence of observed covariates Z, one often considers also the *conditional average treatment effect (CATE)*, which we can define as

$$\mathbb{E}_M(X \mid \mathrm{do}(C = 1), Z) - \mathbb{E}_M(X \mid \mathrm{do}(C = 0), Z).$$

when assuming an underlying iSCM M. There is a large body of literature that considers the question of studying the (asymptotic) efficiency of estimators of the (conditional) average treatment effect. For a nice account of this surprisingly non-trivial inference problem, see e.g. [Wag20].

10.3. Local Causal Discovery

Although the most reliable way to discover causal relations and to estimate their effects is by means of a randomized controlled trial, it is not always possible or feasible to perform such an experiment. One alternative is provided by the Local Causal Discovery (LCD) algorithm [Coo97].

LCD is a *constraint-based* causal discovery algorithm which means that it discovers causal relations by combining the results of conditional independence tests on data. It can be used for the purely observational causal discovery setting where certain background knowledge is available that is weaker than that for the randomized controlled trial. In particular, no randomization is necessary.

The basic idea behind the LCD algorithm is the following result of [Coo97] (originally formulated for L-CBNs, but easily generalized to simple iSCMs):

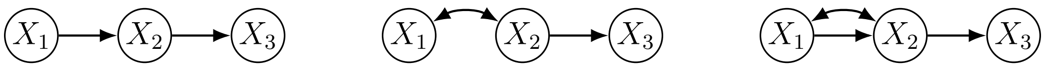


Figure 27: All possible causal graphs detected by LCD.

Proposition 10.3.1. *Let M be a simple SCM with three endogenous variables $V = \{1,2,3\}$ and no exogenous input variables ($J = \emptyset$). Suppose that it is σ-faithful. If X_2 is not a cause of X_1 according to $G(M)$, the following conditional (in)dependencies*[67] *in the observational distribution $P_M(X_1, X_2, X_3)$*

$$X_1 \not\perp\!\!\!\perp X_2, \quad X_2 \not\perp\!\!\!\perp X_3, \quad X_1 \perp\!\!\!\perp X_3 \,|\, X_2$$

imply that the causal graph $G(M)$ must be one of the three DMGs in Figure 27. Hence,

1. *X_3 is not a cause of X_2 according to $G(M)$;*
2. *X_2 is a direct cause of X_3 w.r.t $\{1,2,3\}$ according to $G(M)$;*
3. *X_2 and X_3 are unconfounded according to $G(M)$;*
4. *the causal effect of X_2 on X_3 according to M is identified as:*

$$P_M\big(X_3 \,|\, \mathrm{do}(X_2)\big) = P_M(X_3 \,|\, X_2) \qquad P_M(X_2)\text{-}a.s.. \tag{57}$$

Proof. The proof proceeds by enumerating all (possibly cyclic) DMGs on three variables that the causal graph $G(M)$ could be, and ruling out the ones that do not satisfy the assumptions. The assumption that X_2 is not a cause of X_1 according to G implies that there is no directed edge $X_2 \to X_1$ in the graph $G(M)$. If there were an edge between X_1 and X_3, $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$ would not hold (σ-faithfulness). Also, since $X_1 \not\perp\!\!\!\perp X_2$, X_1 and X_2 must be adjacent (Markov property). Similarly, X_2 and X_3 must be adjacent. X_2 cannot be a collider on any walk between X_1 and X_3 (σ-faithfulness). Since the only possible edges between X_1 and X_2 are $X_1 \to X_2$ and $X_1 \leftrightarrow X_2$ (both of which are into X_2), this means that there must be a directed edge $X_2 \to X_3$, but there cannot be a bidirected edge $X_2 \leftrightarrow X_3$ or directed edge $X_2 \leftarrow X_3$. In other words, the only three possible causal graphs are the ones in Figure 27. The causal do-calculus applied to $G(M)$ yields (57). □

Note that one can apply this to the marginalization $M_{\{1,2,3\}}$ in case $V \supseteq \{1,2,3\}$.

In one of the first applications of LCD, it was discovered that nausea causes vomiting [SC99]. The next example provides another successful application of LCD.

Example 10.3.2. *PIP2 and PIP3 are phospholipids that play an important role in human immune system cells. Figure 28 shows a scatter plot of PIP2 and PIP3 expression levels, measured in individual human immune system cells, after activation of certain*

[67] Henceforth, we will no longer always explicitly write the Markov kernel as a subscript to the conditional independence symbol if it is clear from the context which Markov kernel is meant.

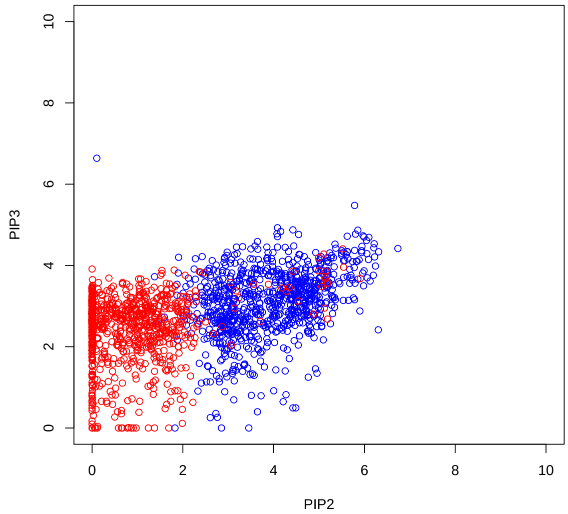


Figure 28: "Text-book" example of an LCD pattern in flow cytometry data of [SPP+05]. See Example 10.3.2 for details.

protein signaling cascades in these cells [SPP+05]. The measurements have been performed under two different experimental conditions: observational ($C = 0$, in blue), and after intervening by administering the chemical compound Psitectorigenin to the cells before measuring the PIP2 and PIP3 levels ($C = 1$, in red). The experimental protocol justifies the assumption that neither PIP2 nor PIP3 expression levels can cause the experimental condition (because these expression levels were measured after the experimental condition had been imposed on the cells). Also, assuming the cells to be properly randomized before split into the two groups (corresponding with the two experimental conditions) we can conclude that the experimental condition C and PIP2 are unconfounded, and similarly that C and PIP3 are unconfounded. The scatter plot suggests the following conditional independence:

$$X_{PIP3} \underset{P(X_{PIP2}, X_{PIP3} \,|\, \mathrm{do}(X_C))}{\perp\!\!\!\perp} X_C \,|\, X_{PIP2},$$

which can also be confirmed with statistical conditional independence tests. Therefore, we have an LCD pattern with $X_1 = X_C$, $X_2 = X_{PIP2}$, $X_3 = X_{PIP3}$, which allows us to infer that the PIP2 expression level causes the PIP3 expression level. Under the randomization assumption, we can even infer that Psitectorigenin exposure is a cause of PIP2 expression levels. This is in line with the Psitectorigenin being known as an inhibitor of PIP2, reducing the quantity of PIP2 in cells after exposure of the cells to this inhibitor.

A high-dimensional adaptation has also been shown to be successful in predicting the effects of gene knockout on gene expression levels from large-scale interventional yeast gene expression data [VM19].

In case more than three variables have been observed, one can run LCD on all triples of variables for which its assumptions apply. In that case, one should keep in mind that a direct edge in a marginalized graph does not imply the presence of the directed edge in the original graph (only the presence of a directed path). In other words, with respect to a larger set of observed variables, the causal relations found by LCD are not necessarily direct.

In case of more than three observed variables, one can also replace the single variable X_2 in the LCD algorithm by a subset of variables, a so-called *separating set*. This idea is exploited efficiently in case of many variables in the Invariant Causal Prediction algorithm [PBM16].

10.4. Y-structures

For both the randomized controlled trial and the LCD algorithm, we need prior knowledge: we need to know already that one of the variables is not a cause of another one. It turns out that in the absence of any such causal background knowledge, we can sometimes still deduce causal relationships from observed conditional independences. The simplest such example is given by the "Y-structure" pattern [Man06]. We here also give the generalization of the Y-structure pattern to simple SCMs.

Proposition 10.4.1. *Let M be a simple SCM with endogenous variables $V = \{1,2,3,4\}$ and no exogenous input variables ($J = \emptyset$). Suppose that it is σ-faithful. The following conditional (in)dependencies in the observational distribution $P_M(X_1, X_2, X_3, X_4)$*

$$\begin{array}{lll} X_1 \not\perp\!\!\!\perp X_4, & X_2 \not\perp\!\!\!\perp X_4, & X_1 \perp\!\!\!\perp X_2, \\ X_1 \perp\!\!\!\perp X_4 \,|\, X_3, & X_2 \perp\!\!\!\perp X_4 \,|\, X_3, & X_1 \not\perp\!\!\!\perp X_2 \,|\, X_3, \end{array}$$

imply that the causal graph $G(M)$ must be one of the nine DMGs in Figure 29. Hence,

1. *X_4 is not a cause of X_3 according to $G(M)$;*
2. *X_3 is a direct cause of X_4 w.r.t. $\{1,2,3,4\}$ according to $G(M)$;*
3. *X_3 and X_4 are unconfounded according to $G(M)$;*
4. *the causal effect of X_3 on X_4 according to M satisfies:*

$$P_M\big(X_4 \,|\, \mathrm{do}(X_3)\big) = P_M(X_4 \,|\, X_3) \qquad P_M(X_3)\text{-a.s.}. \tag{58}$$

Proof. By using the global Markov property and the faithfulness assumption, one can check that the only (cyclic or acyclic) graphs that are compatible with the observed conditional independences are the ones in Figure 29. The statements now follow. □

Note that one can apply this to the marginalization $M_{\{1,2,3,4\}}$ in case $V \supseteq \{1,2,3,4\}$.

This example illustrates how conditional independence patterns in the observational distribution allow one to infer certain features of the underlying causal model. This

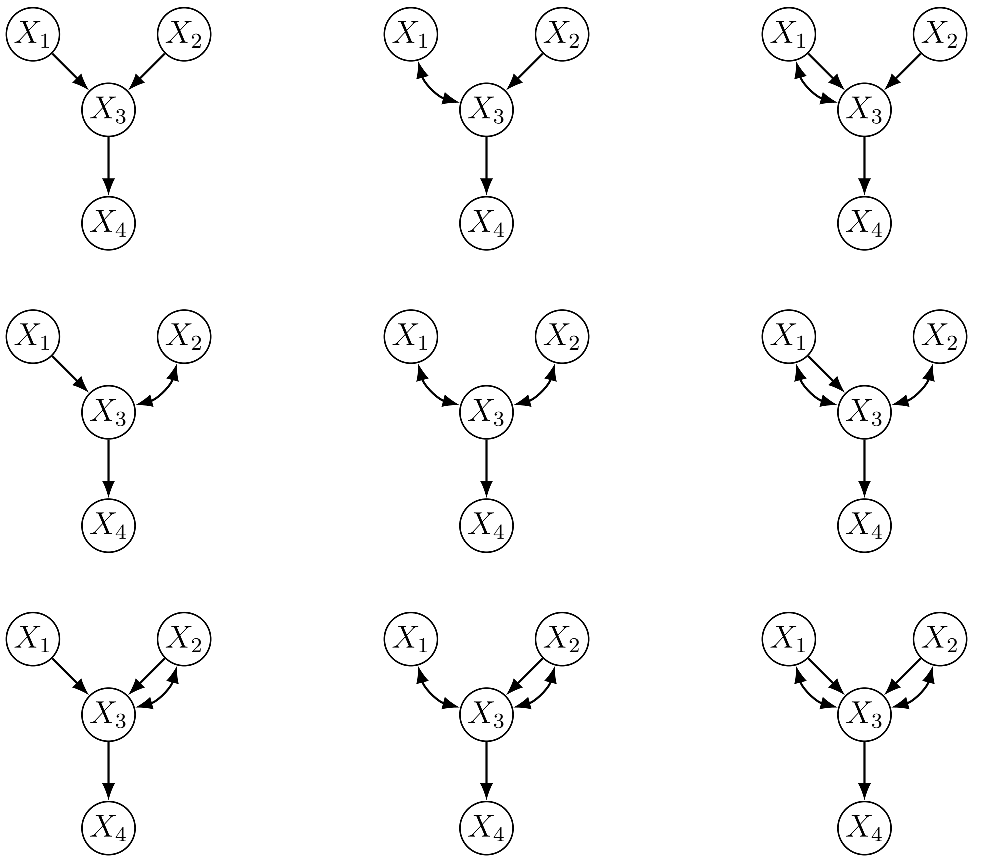


Figure 29: Causal graphs satisfying the "Y-structure" pattern on four variables.

principle is exploited more generally by constraint-based methods, and implicitly, by score-based methods that optimize a penalized likelihood over (equivalence classes of) causal graphs. In Chapter 12 we will describe in detail one of the most sophisticated constraint-based causal discovery methods, Fast Causal Inference (FCI).

Typically, the graph cannot be completely identified from purely observational data. For example, in the Y-structure case, the conditional independences in the observational data do not allow to conclude whether the dependence between X_1 and X_3 is explained by X_1 being a cause of X_3, or by X_1 and X_3 having a latent common cause, or because of some latent selection mechanism, or a combination of those. However, under an appropriate faithfulness assumption, one can deduce the Markov equivalence class of the graph from the conditional independences in the observational data, i.e., the class of all DMGs that induce the same separations.

A significant practical disadvantage of causal discovery methods from purely observational data is that they typically need *very large* sample sizes and *strong assumptions* in order to work reliably (even on simulations).

10.5. Minimal Separating Sets, Minimal Connecting Sets

Minimal separating sets and minimal connecting sets are useful in that they give a relationship between certain separation properties of the graph and ancestral relations in the graph [SMR99, CH11]. This can also be seen as a simple form of causal discovery.

Definition 10.5.1. *Let X, Y, Z, S be sets of nodes in a CDMG G with input nodes J and output nodes V. We say that the* ***minimal σ-separation***

$$X \underset{G}{\overset{\sigma}{\perp}} Y \mid S \cup [Z]$$

holds if and only if

$$X \underset{G}{\overset{\sigma}{\perp}} Y \mid S \cup Z \quad \wedge \quad \forall Q \subsetneq Z : X \underset{G}{\overset{\sigma}{\not\perp}} Y \mid S \cup Q.$$

In words: all nodes in Z are required (in the context of the nodes in S) to σ-separate X from Y. The ***minimal d-separation*** *$X \perp_G^d Y \mid S \cup [Z]$ is defined analogously.*

Minimal separating sets imply the presence of certain ancestral relations (this generalizes a result of [SMR99]). But first we prove a little lemma.

Lemma 10.5.2. *Let G be a CDMG with input nodes J and output nodes V. Let $i, j \in J \cup V$ and $Z \subseteq J \cup V$. If π is a Z-σ-open or Z-d-open walk between i and j in G, then every node on π is in $\mathrm{Anc}_G((\{i, j\} \setminus J) \cup Z)$.*

Proof. Suppose k is a node on π. Then either k is a collider, or there is a directed subwalk from k to a collider on π, or to an endnode of π that is not in J. In all cases, $k \in \mathrm{Anc}_G((\{i, j\} \setminus J) \cup Z)$. This holds for both d-separation and σ-separation. □

Proposition 10.5.3. *Let $\{x\}, \{y\}, S, Z$ be mutually disjoint sets of nodes in a CDMG G with input nodes J and output nodes V. Then:*

$$x \underset{G}{\overset{\sigma}{\perp}} y \mid S \cup [Z] \implies Z \subseteq \mathrm{Anc}_G(\{x, y\} \cup S).$$

A similar statement holds for d-separation.

Proof. Write $A := \mathrm{Anc}_G(\{x, y\} \cup S)$. Let $z \in Z$. Suppose that $z \notin A$. Let $Q = A \cap Z$. Then $z \notin Q$, and therefore $Q \subseteq Z \setminus \{z\}$. Then there is a $(Q \cup S)$-σ-open path π between x and $\{y\} \cup J$ in G. Then every node on π is in $\mathrm{Anc}_G((\{x, y\} \setminus J) \cup Q \cup S)$ (Lemma 10.5.2). Therefore, every node on π is in A. Hence no node in $Z \setminus Q$ is on π. Therefore, adding $(Z \setminus Q)$ to $(Q \cup S)$ cannot σ-block π. Hence $x \not\perp_G^\sigma y \mid Z \cup S$. Contradiction. □

Similarly, we define minimal connections.

Definition 10.5.4. *Let X, Y, Z, S be sets of nodes in a CDMG G with input nodes J and output nodes V. We say that the* ***minimal σ-connection***

$$X \underset{G}{\overset{\sigma}{\not\perp}} Y \mid S \cup [Z]$$

holds if and only if

$$X \underset{G}{\overset{\sigma}{\not\perp}} Y \mid S \cup Z \quad \wedge \quad \forall Q \subsetneq Z : X \underset{G}{\overset{\sigma}{\perp}} Y \mid S \cup Q.$$

In words: all nodes in Z are required (in the context of the nodes in S) to σ-connect X with Y. The ***minimal d-connection*** *$X \not\perp_G^d Y \mid S \cup [Z]$ is defined analogously.*

Note that despite the notation, a minimal connection is *not* the logical negation of a minimal separation.

Minimal connections imply the absence of certain ancestral relations:

Proposition 10.5.5. *Let $\{x\}, \{y\}, S, \{z\}$ be mutually disjoint sets of nodes in a CDMG G with input nodes J and output nodes V. Then*

$$x \underset{G}{\overset{\sigma}{\not\perp}} y \mid S \cup [\{z\}] \implies z \notin \mathrm{Anc}_G(\{x, y\} \cup S)$$

and a similar statement holds for d-separation.

Proof. There exists a $S \cup \{z\}$-σ-open path between x and $y \cup J$ in G that contains a collider in $\mathrm{Anc}_G(\{z\})$ that is not in $\mathrm{Anc}_G(S)$. If $z \in \mathrm{Anc}_G(S)$ this would be a contradiction. If $z \in \mathrm{Anc}_G(x)$, then we can consider the walk between x and y obtained from composing the subpath of the original path between y and the first collider (starting from y) in $\mathrm{Anc}_G(\{z\}) \setminus \mathrm{Anc}_G(S)$ with a directed path to z and then on to x, without passing through nodes in S. This walk between x and y must be σ-open given S, a contradiction. Similarly we obtain a contradiction if $z \in \mathrm{Anc}_G(y)$.

The same proof works also for d-separation. □

11. Independence Testing

In this lecture, we will consider the following questions. How can we test whether...

- ...two random variables are independent?
- ...two random variables are conditionally independent given a third random variable?
- ...a random variable is independent of a non-random variable?
- ...a conditional random variable is conditionally independent of a conditional random variable, given another conditional random variable?

We will consider these questions only for the special case of finite categorical variables, i.e., variables that take values in finite spaces. In particular, we will discuss a test known as the G test. This has been defined in the literature for random variables, but we will extend it here to a general case involving conditional random variables (with "purely" random and "purely" non-random variables as special cases). We will state conditions under which the tests are asymptotically valid and consistent.

11.1. Marginal Independence for Categorical Random Variables

Consider two categorical random variables X, Y taking values in finite spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, with $2 \leq |\mathcal{X}| < \infty$ and $2 \leq |\mathcal{Y}| < \infty$, and joint distribution $P(X, Y)$. We can represent the density in a table (assuming $\mathcal{X} = \{1, \dots, k\}$ and $\mathcal{Y} = \{1, \dots, l\}$):

	$Y = 1$	$Y = 2$	$\dots$	$Y = l$	
$X = 1$	θ_{11}	θ_{12}	$\dots$	θ_{1l}	θ_{1+}
$X = 2$	θ_{21}	θ_{22}	$\dots$	θ_{2l}	θ_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X = k$	θ_{k1}	θ_{k2}	$\dots$	θ_{kl}	θ_{k+}
	θ_{+1}	θ_{+2}	$\dots$	θ_{+l}	$\theta_{++} = 1$

where we introduced the parameter $\theta \in \mathcal{X}_\Theta$ by setting $\theta_{xy} = P(X = x, Y = y)$ for $x \in \mathcal{X}, y \in \mathcal{Y}$. We introduce here the convention that a "+" index denotes summation over that index, i.e.,

$$\theta_{+y} := \sum_{x \in \mathcal{X}} \theta_{xy}, \qquad \theta_{x+} := \sum_{y \in \mathcal{Y}} \theta_{xy}, \qquad \theta_{++} := \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \theta_{xy}.$$

For the parameter space we take the $(|\mathcal{X}||\mathcal{Y}| - 1)$-dimensional simplex:

$$\mathcal{X}_\Theta := \{\theta \in \prod_{(x,y) \in \mathcal{X} \times \mathcal{Y}} [0, 1] : \theta_{++} = 1\}.$$

With Remark 2.5.23, we get:

$$X \underset{P(X,Y)}{\perp\!\!\!\perp} Y \iff P(X,Y) = P(X) \otimes P(Y),$$

where $P(X)$ and $P(Y)$ are the marginal distributions of $P(X,Y)$. In the discrete case we consider here, this holds if and only if

$$\forall x \in \mathcal{X}, y \in \mathcal{Y} : \theta_{xy} = \theta_{x+}\theta_{+y}.$$

The parameters satisfying this constraint form the allowed parameters under the null hypothesis of independence $H_0 : X \perp\!\!\!\perp Y$. We introduce the corresponding restricted parameter space

$$\mathcal{X}_\Theta^0 := \{\theta \in \mathcal{X}_\Theta : \theta_{xy} = \theta_{x+}\theta_{+y} \ \forall x \in \mathcal{X}, y \in \mathcal{Y}\} \subseteq \mathcal{X}_\Theta.$$

We can also write the null hypothesis as $H_0 : \theta \in \mathcal{X}_\Theta^0$. As alternative hypothesis we take that of dependence, i.e., $H_1 : X \not\!\perp\!\!\!\perp Y$, or equivalently, $H_1 : \theta \in \mathcal{X}_\Theta^1$ with $\mathcal{X}_\Theta^1 := \mathcal{X}_\Theta \setminus \mathcal{X}_\Theta^0$.

Suppose now that we have independent and identically distributed data $(X_n, Y_n)_{n=1}^N$ with $(X_n, Y_n) \,|\, \Theta \sim P(X, Y \,|\, \Theta)$ for all $n = 1, \ldots, N$, with the "true" parameter θ unknown. In other words, we assume for the joint distribution on the observed data

$$P((X_n, Y_n)_{n=1}^N \,|\, \Theta) = \bigotimes_{n=1}^N P(X_n, Y_n \,|\, \Theta),$$

where each $P(X_n, Y_n \,|\, \Theta)$ is a copy of the Markov kernel $P(X, Y \,|\, \Theta)$. We define the *counts* as the number of observations with a given value $(x, y) \in \mathcal{X} \times \mathcal{Y}$:

$$N_{xy} := \sum_{n=1}^N \mathbb{1}_{(x,y)}(X_n, Y_n).$$

We can represent them in a contingency table:

	$Y=1$	$Y=2$	$\ldots$	$Y=l$	
$X=1$	N_{11}	N_{12}	$\ldots$	N_{1l}	N_{1+}
$X=2$	N_{21}	N_{22}	$\ldots$	N_{2l}	N_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X=k$	N_{k1}	N_{k2}	$\ldots$	N_{kl}	N_{k+}
	N_{+1}	N_{+2}	$\ldots$	N_{+l}	$N_{++} = N$

where we used a similar summation convention for the counts as for the parameters.

The classical frequentist procedure for deciding between two hypotheses $H_0 : \theta \in \mathcal{X}_\Theta^0$ and $H_1 : \theta \in \mathcal{X}_\Theta \setminus \mathcal{X}_\Theta^0$ is as follows. One comes up with a *test statistic* $T(D)$, which is a function of the data $D \sim P(D \,|\, \Theta)$, whose value should help us distinguish between the two hypotheses. We will consider here one-sided tests, where a large value of $T(D)$ is in favor of H_1 while a small value of $T(D)$ is in favor of H_0. Then one chooses a particular significance level $\alpha \in (0,1)$. From the observed data d, one then calculates a corresponding p-value $p(d)$, which is the probability under the null hypothesis that the test statistic has the observed or a more extreme value. For the one-sided tests we will consider here, the p-value can be defined as

$$p(d) := \sup_{\theta \in \mathcal{X}_\Theta^0} P(T(D) \geq T(d) \,|\, \theta).$$

Then, a decision is taken: if $p(d) \leq \alpha$, one considers this as sufficient evidence to reject H_0 (and accept H_1), while if $p(d) > \alpha$, one does not reject H_0 as the evidence in the data is considered insufficient to do so. Often, the main desideratum is to control the probability of a Type I error (i.e., the error of incorrectly rejecting the null hypothesis), which can be achieved by choosing α sufficiently small. Indeed, from the definition of the p-value it follows that:

$$\forall \alpha \in (0,1)\, \forall \theta \in \mathcal{X}_\Theta^0 : \; P(p(D) \leq \alpha \,|\, \theta) \leq \alpha.$$

For causal discovery, however, we need a more symmetric treatment of the two hypotheses, as there we require both the probability of a Type I error and of a Type II error (i.e., the error of incorrectly rejecting the alternative hypothesis) to be small. Before we investigate this tradeoff, let us first propose a concrete test statistic for the case at hand and obtain an approximate expression for the corresponding p-value.

Here we will work out the details of the *likelihood ratio test*, which for this particular case is also known as the *G test*. We start by writing down the likelihood:

$$\theta \mapsto P((X_n, Y_n)_{n=1}^N \,|\, \theta) = \prod_{n=1}^N \theta_{X_n Y_n} = \prod_{x \in \mathcal{X}} \prod_{y \in \mathcal{Y}} \theta_{xy}^{N_{xy}}$$

where we used the counts as a sufficient statistic of the data. This is a multinomial distribution with parameters $(\theta_{xy})_{x \in \mathcal{X}, y \in \mathcal{Y}}$ and N. Maximizing the likelihood with respect to the parameters $\theta \in \mathcal{X}_\Theta$, we obtain the well-known maximum likelihood estimator[68]

$$\hat{\theta}_{xy} = \frac{N_{xy}}{N},$$

i.e., the fractions of the different outcomes in the data. Under the null hypothesis H_0, $\theta_{xy} = \theta_{x+}\theta_{+y}$, and the likelihood factorizes:

$$\begin{aligned} \theta \in \mathcal{X}_\Theta^0 \implies P((X_n, Y_n)_{n=1}^N \,|\, \theta) &= \prod_{x \in \mathcal{X}} \prod_{y \in \mathcal{Y}} (\theta_{x+}\theta_{+y})^{N_{xy}} \\ &= \left(\prod_{x \in \mathcal{X}} \prod_{y \in \mathcal{Y}} \theta_{x+}^{N_{xy}} \right) \left(\prod_{x \in \mathcal{X}} \prod_{y \in \mathcal{Y}} \theta_{+y}^{N_{xy}} \right) \\ &= \left(\prod_{x \in \mathcal{X}} \theta_{x+}^{N_{x+}} \right) \left(\prod_{y \in \mathcal{Y}} \theta_{+y}^{N_{+y}} \right). \end{aligned}$$

[68]We follow the notational convention in statistics by writing estimators in lowercase (e.g., $\hat{\theta}$ rather than $\hat{\Theta}$), even though they are random variables.

This is just the product of two independent multinomial distributions with (variationally independent) parameters $(\theta_{x+})_{x\in\mathcal{X}}$ and $(\theta_{+y})_{y\in\mathcal{Y}}$ (and N), respectively. Hence, the restricted maximum likelihood estimator under H_0 is

$$\hat{\theta}^0_{xy} = \hat{\theta}^0_{x+}\hat{\theta}^0_{+y} = \frac{N_{x+}}{N}\frac{N_{+y}}{N}.$$

The likelihood ratio is obtained by dividing the likelihood for $\hat{\theta}$ by the likelihood for $\hat{\theta}^0$:

$$\begin{aligned}\frac{\sup_{\theta\in\mathcal{X}^0_\Theta\cup\mathcal{X}^1_\Theta} P((X_n,Y_n)^N_{n=1}\,|\,\theta)}{\sup_{\theta\in\mathcal{X}^0_\Theta} P((X_n,Y_n)^N_{n=1}\,|\,\theta)} &= \frac{P((X_n,Y_n)^N_{n=1}\,|\,\hat{\theta})}{P((X_n,Y_n)^N_{n=1}\,|\,\hat{\theta}^0)} = \prod_{x\in\mathcal{X}}\prod_{y\in\mathcal{Y}}\left(\frac{\hat{\theta}_{xy}}{\hat{\theta}^0_{x+}\hat{\theta}^0_{+y}}\right)^{N_{xy}} \\ &= \prod_{x\in\mathcal{X}}\prod_{y\in\mathcal{Y}}\left(\frac{N_{xy}N}{N_{x+}N_{+y}}\right)^{N_{xy}}.\end{aligned} \tag{59}$$

The likelihood ratio test statistic is defined as 2 times the natural logarithm of this ratio:

$$G_N := 2\log\frac{\sup_{\theta\in\mathcal{X}^0_\Theta\cup\mathcal{X}^1_\Theta} P((X_n,Y_n)^N_{n=1}\,|\,\theta)}{\sup_{\theta\in\mathcal{X}^0_\Theta} P((X_n,Y_n)^N_{n=1}\,|\,\theta)} = 2\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} N_{xy}\log\frac{N_{xy}N}{N_{x+}N_{+y}}. \tag{60}$$

Since counts can be zero, one should interpret $0\log\frac{0}{n}$ in this expression as 0 (for $n\in\mathbb{N}$).

We will now consider the asymptotic behavior of the test statistic under the null hypothesis. This will yield an approximation for the p-value that we can use also for finite samples. As a simplifying assumption, we will henceforth assume that all probabilities are positive,[69] i.e.,

$$\theta_{xy} > 0 \qquad \forall x\in\mathcal{X}, y\in\mathcal{Y}. \tag{61}$$

Proposition 11.1.1. *Under $H_0: X \perp\!\!\!\perp Y$, and with regularity assumption (61),*

$$G_N \rightsquigarrow \chi^2_\nu$$

with $\nu = (|\mathcal{X}|-1)(|\mathcal{Y}|-1)$ as sample size $N\to\infty$. In words, the test statistic G_N converges in distribution[70] as $N\to\infty$ to a chi-squared distribution with ν degrees of freedom.[71]

Proof. This is a direct application of Theorem 4.43 in [BJvdV17], for which a proof is provided in Chapter 16 in [vdV98]. One has to be careful here to use a different parameterization—in terms of (variationally) independent parameters—i.e., such

[69] The singularities for vanishing values of θ_{xy} can be dealt with, but require special attention. For simplicity we study only the regular case here.

[70] We say that a sequence of real-valued random variables $X_1, X_2, \ldots$ *converges in distribution* to X_∞, and write $X_n \rightsquigarrow X_\infty$, if $P(X_n \le x)\to P(X_\infty\le x)$ for all $x\in\mathbb{R}$ such that $\xi\mapsto P(X_\infty\le\xi)$ is continuous at x.

[71] The chi-square distribution with ν degrees of freedom is defined as the distribution of a sum of squares of ν independent standard normal random variables, i.e, of $\sum_{i=1}^{\nu} Z_i^2$ where $Z_i\sim N(0,1)$ are i.i.d..

that the parameter space contains an open part of $\mathbb{R}^{|\mathcal{X}||\mathcal{Y}|-1}$, when calculating the score function and the Fisher information matrix when checking the regularity conditions. For example, one can choose a pair $(k,l) \in \mathcal{X} \times \mathcal{Y}$ and take parameters $\theta_{x,y} = \vartheta_{x,y}$ for $x \neq k$ or $y \neq l$, and $\theta_{k,l} = 1 - \sum_{(x,y)\neq(k,l)} \vartheta_{x,y}$. The dimensionality of $\mathcal{X}_\Theta$ is $|\mathcal{X}||\mathcal{Y}| - 1$, while that of $\mathcal{X}_\Theta^0$ is $(|\mathcal{X}| - 1) + (|\mathcal{Y}| - 1)$. The degrees of freedom of the asymptotic chi-square distribution is the difference of the two, i.e., $\nu = (|\mathcal{X}||\mathcal{Y}| - 1) - ((|\mathcal{X}| - 1) + (|\mathcal{Y}| - 1)) = (|\mathcal{X}| - 1)(|\mathcal{Y}| - 1)$.

An alternative proof will be provided later in a more general setting (see Proposition 11.3.2). □

One therefore obtains an approximate level α test (i.e., a test with Type I error asymptotically upper bounded by α) by rejecting H_0 when $G_N \geq \chi^2_{\nu,1-\alpha}$. Here, $\chi^2_{\nu,1-\alpha} := F^{-1}_{\chi^2_\nu}(1-\alpha)$ is the upper α quantile of the χ^2-distribution with ν degrees of freedom, with $F_{\chi^2_\nu}$ the corresponding distribution function (cumulative density function) and $F^{-1}_{\chi^2_\nu}$ its inverse (i.e., the quantile function). Indeed, if $\theta \in \mathcal{X}_\Theta^0$, then $P(G_N \geq \chi^2_{\nu,1-\alpha} \,|\, \theta) \to \alpha$, for any $\alpha \in (0,1)$. Since

$$G_N \geq \chi^2_{\nu,1-\alpha} \iff G_N \geq F^{-1}_{\chi^2_\nu}(1-\alpha) \iff F_{\chi^2_\nu}(G_N) \geq 1-\alpha \iff 1 - F_{\chi^2_\nu}(G_N) \leq \alpha,$$

the corresponding approximate p-value is $1 - F_{\chi^2_\nu}(G_N)$; if this is smaller than or equal to the chosen threshold α, we reject H_0. This test is called the *G-test*.

But what about the Type II error? If we let the sample size N grow, we would hope that the probability of a wrong test result becomes arbitrarily small, and vanishes in the limit $N \to \infty$.

Definition 11.1.2. *A (conditional) independence test is called **consistent** if the probabilities of both Type I and Type II errors converge to 0, no matter what the true parameter value is.*

To obtain consistency, it is not an option to just control Type I error at a fixed level α; instead, one has to use a level α_N that depends on the sample size N, and converges to 0 (implying that Type I error converges to 0). However, because of the tradeoff between Type I and Type II errors, the rate at which α_N converges to 0 has to be chosen carefully in order to be able to guarantee that also Type II error vanishes asymptotically. As we shall see, the convergence rate of α_N should be chosen sufficiently slow.

While it is often easier to calculate the Type I error than the Type II error of a test, in this case we can actually analyze the asymptotic behavior of the test statistic under the alternative hypothesis H_1. Define

$$\hat{I}_N := \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \hat{\theta}_{xy} \log \frac{\hat{\theta}_{xy}}{\hat{\theta}_{x+}\hat{\theta}_{+y}} = \frac{G_N}{2N}$$

where we used that $\hat{\theta}^0_{x+} = \hat{\theta}_{x+}$ and $\hat{\theta}^0_{+y} = \hat{\theta}_{+y}$. This is an estimator (the so-called "plug-in estimator" $I(\hat{\theta})$) of the mutual information $I(X;Y)$:

$$\begin{aligned} I(\theta) &:= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \theta_{xy}\log\frac{\theta_{xy}}{\theta_{x+}\theta_{+y}} \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(X=x,Y=y\,|\,\theta)\log\frac{P(X=x,Y=y\,|\,\theta)}{P(X=x\,|\,\theta)P(Y=y\,|\,\theta)} =: I(X;Y\,\|\,\theta). \end{aligned}$$

With Jensen's inequality, one can show that $I(X;Y\,\|\,\theta)\geq 0$, and that $I(X;Y\,\|\,\theta) = 0 \iff X \perp\!\!\!\perp Y\,|\,\Theta=\theta$. Note further that the function $\mathcal{X}_\Theta \to [0,\infty) : \theta \mapsto I(\theta)$ is continuous.

With this observation, we can prove the asymptotic consistency of the G-test under assumptions on the critical values used for deciding between H_0 and H_1.

Corollary 11.1.3. *Consider an infinite sequence of G tests performed on the first N samples of an infinitely large data set $(X_n,Y_n)_{n=1}^\infty$, where one accepts $H_1 : X \not\perp\!\!\!\perp Y$ if $G_N \geq \tau_N$, and otherwise accepts $H_0 : X \perp\!\!\!\perp Y$, for some given sequence of thresholds τ_N. Under the regularity assumption (61), this sequence of tests is asymptotically consistent if $\tau_N\to\infty$ but $\tau_N/N\to 0$.*

Proof. We start by a simple application of the strong law of large numbers. Let $\theta\in\mathcal{X}_\Theta$. Since the (X_n,Y_n) are assumed to be i.i.d., and

$$\mathbb{E}\big(\mathbb{1}_{(x,y)}(X_n,Y_n)\big) = \theta_{xy}$$

for all $x\in\mathcal{X}, y\in\mathcal{Y}$, we conclude that $N_{xy}/N \overset{a.s.}{\to} \theta_{xy}$ for all $x\in\mathcal{X}, y\in\mathcal{Y}$ by the strong law of large numbers.[72] Hence $\hat{\theta}_{xy} \overset{a.s.}{\to} \theta_{xy}$. Hence, also $\hat{\theta}_{+x} \overset{a.s.}{\to} \theta_{+x}$ and $\hat{\theta}_{y+} \overset{a.s.}{\to} \theta_{y+}$. Furthermore, by continuity, $I(\hat{\theta}) \overset{a.s.}{\to} I(\theta)$. Hence, $G_N/N \overset{a.s.}{\to} 2I(\theta)$.

Under H_1, we have $I(\theta)>0$, and since by assumption $\tau_N/N\to 0$, $\mathbb{1}_{G_N<\tau_N} \overset{a.s.}{\to} 0$. Since a.s. convergence implies convergence in probability,

$$\theta\in\mathcal{X}_\Theta^1 \implies P(G_N<\tau_N\,|\,\theta)\to 0.$$

Thus, the probability of a Type II error vanishes asymptotically.

This same approach doesn't work for the Type I error. The reason is that even though we assume $\tau_N\to\infty$, and we know that $G_N/N \overset{a.s.}{\to} 0$ under H_0, this does not suffice to conclude anything about the probability of the event $G_N\geq\tau_N$. But we can make use of Proposition 11.1.1, which states that $G_N \rightsquigarrow \chi^2_\nu$ under H_0. Since the distribution function of χ^2_ν is continuous, this implies uniform convergence of the distribution functions:

$$\sup_{x\in\mathbb{R}} |F_{G_N}(x) - F_{\chi^2_\nu}(x)| \to 0.$$

Hence

$$|F_{G_N}(\tau_N) - F_{\chi^2_\nu}(\tau_N)| \leq \sup_{x\in\mathbb{R}} |F_{G_N}(x) - F_{\chi^2_\nu}(x)| \to 0.$$

[72] The convergence is "almost surely", i.e., $N_{xy}/N \overset{a.s.}{\to} \theta_{xy}$ means that $P(N_{xy}/N\to\theta_{xy}\,|\,\theta)=1$.

Since $\tau_N \to \infty$, $F_{\chi^2_\nu}(\tau_N) \to 1$. Hence, also $F_{G_N}(\tau_N) \to 1$. We conclude that

$$\theta \in \mathcal{X}^0_\Theta \implies P(G_N \geq \tau_N \,|\, \theta) \to 0,$$

i.e., the probability of a Type I error converges to 0. □

While one traditionally focuses mostly on Type I error control, in causal discovery we are more interested in having both small Type I and Type II error. In order to achieve this (at least asymptotically, i.e., for sufficiently large sample sizes), we can thus make use of a sequence of thresholds that satisfies the assumptions in the corollary. In terms of p-values, this means that to bound the Type I error, a fixed critical value α suffices, but for consistency we let $\alpha_N \to 0$ with a rate such that $\chi^2_{\nu,1-\alpha_N}/N \to 0$.

While for a finite sample, we can give guarantees (at least approximately) on the Type I error, it will often be impossible to provide guarantees on the Type II error without making strong assumptions on the parameters. Indeed, since the mutual information $I(X;Y)$ (a measure of the dependence of X and Y) can be arbitrarily close to zero for weakly dependent X and Y, one cannot know in advance how many samples will be needed to be able to distinguish it from an independence.[73]

11.2. Conditional Independence for Categorical Random Variables

We now extend the G test to a conditional independence test that we will refer to as the conditional G test.

Consider three categorical random variables X, Y, Z taking values in spaces $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, respectively (with $2 \leq |\mathcal{X}| < \infty$, $2 \leq |\mathcal{Y}| < \infty$ and $1 \leq |\mathcal{Z}| < \infty$) and joint distribution $P(X, Y, Z)$. With Remark 2.5.23, we get (because finite spaces are standard):

$$\begin{aligned}
X \underset{P(X,Y,Z)}{\perp\!\!\!\perp} Y \,|\, Z &\iff P(X,Y,Z) = P(X|Z) \otimes P(Y,Z) \\
&\iff P(X,Y|Z) = P(X|Z) \otimes P(Y|Z) \quad P(Z)\text{-a.s.} \\
&\iff \forall z \in \mathcal{Z} : \big[P(Z=z) > 0 \implies \\
&\qquad\qquad P(X,Y \,|\, Z=z) = P(X \,|\, Z=z)P(Y \,|\, Z=z)\big] \\
&\iff \forall z \in \mathcal{Z} : \big[P(Z=z) > 0 \implies X \underset{P(X,Y \,|\, Z=z)}{\perp\!\!\!\perp} Y\big].
\end{aligned}$$

This suggests that we can make use of an independence test for two categorical variables on each "stratum" corresponding to conditioning on a specific value $Z = z$ that has positive probability to occur.

We parameterize the conditional kernel $P(X, Y|Z)$ in terms of parameters $(\theta_{xy|z})_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}}$ which live in space

$$\mathcal{X}_{\Theta_{XY|Z}} := \{\theta \in \prod_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}} [0,1] : \forall_{z\in\mathcal{Z}} \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \theta_{xy|z} = 1\}.$$

[73] This is referred to as the lack of "uniformly consistent" (conditional) independence tests.

With the summation convention, we can write the normalization condition as $\theta_{++|z} = 1$ for all $z \in \mathcal{Z}$. For those $z \in \mathcal{Z}$ with $P(Z = z) > 0$, we have

$$\frac{P(X = x, Y = y, Z = z \mid \Theta = \theta)}{P(Z = z \mid \Theta = \theta)} = P(X = x, Y = y \mid Z = z, \Theta = \theta) = \theta_{xy|z}.$$

We also parameterize the marginal distribution $P(Z)$ in terms of parameters $(\theta_z)_{z \in \mathcal{Z}}$ which live in space

$$\mathcal{X}_{\Theta_Z} := \{\theta \in \prod_{z \in \mathcal{Z}} [0, 1] : \sum_{z \in \mathcal{Z}} \theta_z = 1\}.$$

Any joint distribution of X, Y and Z can then be parameterized as

$$P(X = x, Y = y, Z = z \mid \Theta = \theta) = \theta_z \theta_{xy|z},$$

with parameter space

$$\mathcal{X}_\Theta := \mathcal{X}_{\Theta_Z} \times \mathcal{X}_{\Theta_{XY|Z}}.$$

We formulate the null hypothesis $H_0 : X \perp\!\!\!\perp Y \mid Z$ of independence in terms of the parameters as

$$\forall x \in \mathcal{X}, y \in \mathcal{Y}, \forall z \in \mathcal{Z} : \quad \theta_{xy|z} = \theta_{x+|z} \theta_{+y|z}$$

(for convenience, we have strengthened it a bit; strictly speaking, we only need this relation to hold for all $z \in \mathcal{Z}$ with $\theta_z > 0$; however, since the data will not convey any information on $\theta_{xy|z}$ for such z, this does not matter). The corresponding restricted parameter space is

$$\mathcal{X}^0_{\Theta_{XY|Z}} := \{\theta \in \mathcal{X}_\Theta : \theta_{xy|z} = \theta_{x+|z} \theta_{+y|z} \ \forall x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}\}.$$

We can then also write the null hypothesis as $H_0 : \theta \in \mathcal{X}^0_\Theta$ with $\mathcal{X}^0_\Theta := \mathcal{X}_{\Theta_Z} \times \mathcal{X}^0_{\Theta_{XY \mid Z}}$. As alternative hypothesis we take that of dependence, i.e., $H_1 : X \not\perp\!\!\!\perp Y \mid Z$, or equivalently, $H_1 : \theta \in \mathcal{X}^1_\Theta$, where $\mathcal{X}^1_\Theta := \mathcal{X}_{\Theta_Z} \times \mathcal{X}^1_{\Theta_{XY \mid Z}}$ with $\mathcal{X}^1_{\Theta_{XY \mid Z}} := \mathcal{X}_{\Theta_{XY \mid Z}} \setminus \mathcal{X}^0_{\Theta_{XY \mid Z}}$.

Suppose now that we have independent and identically distributed data $(X_n, Y_n, Z_n)_{n=1}^N$ with $(X_n, Y_n, Z_n) \sim P(X, Y, Z \mid \Theta)$ for all $n = 1, \dots, N$, with the "true" parameter $\theta \in \mathcal{X}_\Theta$ unknown. In other words, we assume for the joint distribution on the observed data

$$P((X_n, Y_n, Z_n)_{n=1}^N \mid \Theta) = \bigotimes_{n=1}^N P(X_n, Y_n, Z_n \mid \Theta),$$

where each $P(X_n, Y_n, Z_n \mid \Theta)$ is a copy of the Markov kernel $P(X, Y, Z \mid \Theta)$. We define the *counts* as the number of observations with a given value $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$:

$$N_{xyz} := \sum_{n=1}^N \mathbb{1}_{(x,y,z)}(X_n, Y_n, Z_n).$$

We again work out the details of the likelihood ratio test, and start by writing down the likelihood:

$$\theta \mapsto P((X_n, Y_n, Z_n)_{n=1}^N \,|\, \theta) = \prod_{n=1}^{N} (\theta_{X_n,Y_n|Z_n} \theta_{Z_n}) = \prod_{x\in\mathcal{X}} \prod_{y\in\mathcal{Y}} \prod_{z\in\mathcal{Z}} (\theta_{xy|z}^{N_{xyz}} \theta_z^{N_{xyz}})$$
$$= \left(\prod_{z\in\mathcal{Z}} \theta_z^{N_{++z}}\right) \left(\prod_{z\in\mathcal{Z}} \prod_{x\in\mathcal{X}} \prod_{y\in\mathcal{Y}} \theta_{xy|z}^{N_{xyz}}\right)$$

where we used the counts as a sufficient statistic of the data. We recognize the first factor as the likelihood of a multinomial distribution with parameters $(\theta_z)_{z\in\mathcal{Z}}$ and N. The second factor is a product of the likelihoods of multinomial distributions with parameters $(\theta_{xy|z})_{x\in\mathcal{X},y\in\mathcal{Y}}$ and N_{++z}, for each $z \in \mathcal{Z}$. Maximizing the likelihood with respect to the parameters $\theta \in \mathcal{X}_\Theta$, we obtain the maximum likelihood estimator

$$\left(\hat{\theta}_{xy|z}, \hat{\theta}_z\right) = \left(\frac{N_{xyz}}{N_{++z}}, \frac{N_{++z}}{N}\right).$$

Under the null hypothesis H_0, $\theta_{xy|z} = \theta_{x+|z}\theta_{+y|z}$, and the likelihood factorizes over X and Y:

$$\theta \in \mathcal{X}_\Theta^0 \implies P((X_n, Y_n, Z_n)_{n=1}^N \,|\, \theta) = \left(\prod_{z\in\mathcal{Z}} \theta_z^{N_{++z}}\right) \prod_{z\in\mathcal{Z}} \prod_{x\in\mathcal{X}} \prod_{y\in\mathcal{Y}} (\theta_{x+|z}\theta_{+y|z})^{N_{xyz}}$$
$$= \left(\prod_{z\in\mathcal{Z}} \theta_z^{N_{++z}}\right) \prod_{z\in\mathcal{Z}} \left(\prod_{x\in\mathcal{X}} \theta_{x+|z}^{N_{x+z}}\right) \left(\prod_{y\in\mathcal{Y}} \theta_{+y|z}^{N_{+yz}}\right).$$

The restricted maximum likelihood estimator under H_0 is

$$\left(\hat{\theta}^0_{xy|z}, \hat{\theta}^0_z\right) = \left(\hat{\theta}^0_{x+|z}\hat{\theta}^0_{+y|z}, \hat{\theta}^0_z\right) = \left(\frac{N_{x+z}N_{+yz}}{N_{++z}^2}, \frac{N_{++z}}{N}\right).$$

The likelihood ratio is obtained by dividing the likelihood for $\hat{\theta}$ by the likelihood for $\hat{\theta}^0$:

$$\frac{\sup_{\theta\in\mathcal{X}_\Theta^0\cup\mathcal{X}_\Theta^1} P((X_n, Y_n, Z_n)_{n=1}^N \,|\, \theta)}{\sup_{\theta\in\mathcal{X}_\Theta^0} P((X_n, Y_n, Z_n)_{n=1}^N \,|\, \theta)} = \frac{P((X_n, Y_n, Z_n)_{n=1}^N \,|\, \hat{\theta})}{P((X_n, Y_n, Z_n)_{n=1}^N \,|\, \hat{\theta}^0)} = \prod_{z\in\mathcal{Z}} \prod_{x\in\mathcal{X}} \prod_{y\in\mathcal{Y}} \left(\frac{\hat{\theta}_{xy|z}}{\hat{\theta}^0_{x+|z}\hat{\theta}^0_{+y|z}}\right)^{N_{xyz}}$$
$$= \prod_{z\in\mathcal{Z}} \prod_{x\in\mathcal{X}} \prod_{y\in\mathcal{Y}} \left(\frac{N_{xyz}N_{++z}}{N_{x+z}N_{+yz}}\right)^{N_{xyz}},$$

where the factors involving the marginal $P(Z)$ cancel out. The likelihood ratio test statistic is defined as 2 times the natural logarithm of this ratio:

$$G_N := 2\log \frac{\sup_{\theta\in\mathcal{X}\Theta^0\cup\mathcal{X}_\Theta^1} P((X_n, Y_n, Z_n)_{n=1}^N \,|\, \theta)}{\sup_{\theta\in\mathcal{X}_\Theta^0} P((X_n, Y_n, Z_n)_{n=1}^N \,|\, \theta)} = 2\sum_{z\in\mathcal{Z}} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} N_{xyz} \log \frac{N_{xyz}N_{++z}}{N_{x+z}N_{+yz}}. \tag{62}$$

We will now consider the asymptotic behavior of the test statistic under both hypotheses. As a simplifying assumption, we will henceforth assume that all probabilities are positive, i.e.,

$$\begin{cases} \theta_z > 0 & \forall z \in \mathcal{Z}, \text{ and} \\ \theta_{xy|z} > 0 & \forall x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}. \end{cases} \tag{63}$$

Proposition 11.2.1. *Under $H_0 : X \perp\!\!\!\perp Y \mid Z$, and with regularity assumption (63)*

$$G_N \rightsquigarrow \chi^2_\nu$$

with $\nu = |\mathcal{Z}|(|\mathcal{X}| - 1)(|\mathcal{Y}| - 1)$ as sample size $N \to \infty$. In words, the test statistic G_N converges in distribution as $N \to \infty$ to a chi-squared distribution with ν degrees of freedom.

Proof. This is analogous to the proof of Proposition 11.1.1. The dimensionality of $\mathcal{X}_\Theta$ is $|\mathcal{Z}|(|\mathcal{X}||\mathcal{Y}| - 1) + (|\mathcal{Z}| - 1)$, while that of $\mathcal{X}^0_\Theta$ is $|\mathcal{Z}|\big((|\mathcal{X}| - 1) + (|\mathcal{Y}| - 1)\big) + (|\mathcal{Z}| - 1)$. The degrees of freedom of the asymptotic chi-square distribution is the difference of the two, i.e., $\nu = |\mathcal{Z}|(|\mathcal{X}| - 1)(|\mathcal{Y}| - 1)$. □

One therefore obtains an approximate level α test (i.e., a test with Type I error asymptotically upper bounded by α) by rejecting H_0 when $G_N \geq \chi^2_{\nu,1-\alpha}$.

Define

$$\hat{I}_N := \sum_{z \in \mathcal{Z}} \hat{\theta}_z \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \hat{\theta}_{xy|z} \log \frac{\hat{\theta}_{xy|z}}{\hat{\theta}_{x+|z} \hat{\theta}_{+y|z}} = \frac{G_N}{2N}$$

where we used that $\hat{\theta}^0_{x+|z} = \hat{\theta}_{x+|z}$ and $\hat{\theta}^0_{+y|z} = \hat{\theta}_{+y|z}$. This is a plug-in estimator of the conditional mutual information $I(X;Y|Z)$:

$$\begin{aligned} I(\theta) &:= \sum_{z \in \mathcal{Z}} \theta_z \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \theta_{xy|z} \log \frac{\theta_{xy|z}}{\theta_{x+|z} \theta_{+y|z}} \\ &= \sum_{z \in \mathcal{Z}} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(X = x, Y = y, Z = z \mid \Theta = \theta) \log \frac{P(X = x, Y = y|Z = z, \Theta = \theta)}{P(X = x|Z = z, \Theta = \theta) P(Y = y|Z = z, \Theta = \theta)} \\ &=: I(X;Y|Z \,\|\, \theta). \end{aligned}$$

With Jensen's inequality, one can show that $I(X;Y|Z) \geq 0$, and that $I(X;Y|Z) = 0 \iff X \perp\!\!\!\perp Y \mid Z$. Note further that the function $\mathcal{X}_\Theta \to [0, \infty) : \theta \mapsto I(\theta)$ is continuous.

With this observation, we can prove the asymptotic consistency of the conditional G-test under assumptions on the critical values used for deciding between H_0 and H_1.

Corollary 11.2.2. *Consider an infinite sequence of conditional G tests performed on the first N samples of an infinitely large data set $(X_n, Y_n, Z_n)_{n=1}^\infty$, where one accepts $H_1 : X \not\perp\!\!\!\perp Y \mid Z$ if $G_N \geq \tau_N$, and otherwise accepts $H_0 : X \perp\!\!\!\perp Y \mid Z$, for some given sequence of thresholds τ_N. Under the regularity assumption (63), this sequence is asymptotically consistent if $\tau_N \to \infty$ but $\tau_N / N \to 0$.*

Proof. This is very similar to the proof of Corollary 11.1.3.

We again apply the strong law of large numbers. Let $\theta \in \mathcal{X}_\Theta$. Since (X_n, Y_n, Z_n) are assumed to be i.i.d., and

$$\mathbb{E}\big(\mathbb{1}_{(x,y,z)}(X_n, Y_n, Z_n)\big) = \theta_{xy|z}\theta_z$$

for all $x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}$, we conclude that $N_{xyz}/N \overset{a.s.}{\to} \theta_{xy|z}\theta_z$ for all $x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}$ by the strong law of large numbers. Hence, also $N_{++z}/N \overset{a.s.}{\to} \theta_z$ for all $z \in \mathcal{Z}$. Hence, using (63), $\hat{\theta}_z \overset{a.s.}{\to} \theta_z$, $\hat{\theta}_{xy|z} \overset{a.s.}{\to} \theta_{xy|z}$, $\hat{\theta}_{+x|z} \overset{a.s.}{\to} \theta_{+x|z}$, and $\hat{\theta}_{y+|z} \overset{a.s.}{\to} \theta_{y+|z}$. By continuity, $I(\hat{\theta}) \overset{a.s.}{\to} I(\theta)$. Hence, $G_N/N \overset{a.s.}{\to} 2I(\theta)$.

We can now reason analogously as in the proof of Corollary 11.1.3 to conclude that the probability of a Type II error vanishes asymptotically.

For an asymptotic estimate of the probability of a Type I error, we can make use of Proposition 11.2.1, which states that $G_N \rightsquigarrow \chi^2_\nu$. This part of the proof is identical to the corresponding part of the proof of Corollary 11.1.3. □

11.3. Marginal Independence of a Random and a Non-Random Variable

Consider two variables X, C taking values in finite spaces $\mathcal{X}, \mathcal{C}$, respectively, (i.e., with $2 \leq |\mathcal{X}| < \infty$ and $2 \leq |\mathcal{C}| < \infty$). Assume that X is a random variable, while C is an input variable, and consider a Markov kernel $K(X \mid C) : \mathcal{C} \dashrightarrow \mathcal{X}$. We will derive a test for the independence

$$X \underset{K(X|C)}{\perp\!\!\!\perp} C.$$

With Definition 2.5.17, this independence holds if and only if there exists a Markov kernel $Q(X) : * \dashrightarrow \mathcal{X}$ such that:

$$K(X \mid C) = Q(X).$$

The latter means that

$$K(X \mid C = c) = Q(X) \tag{64}$$

for all $c \in \mathcal{C}$.

Suppose we obtain data $(X_n, c_n)_{n=1}^N$ such that the X_n are conditionally independent and identically distributed given c_n for all $n = 1, \dots, N$. In other words, we assume the data is sampled from the following Markov kernel:

$$K((X_n)_{n=1}^N \mid (c_n)_{n=1}^N) = \bigotimes_{n=1}^N K(X_n \mid C_n = c_n),$$

where each $K(X_n \mid C_n)$ is a copy of the ("true" but unknown kernel) $K(X \mid C)$.

Note that this is a weaker assumption regarding the sampling scheme than we would have made if C were random. In particular, we make no assumption at all regarding how the sequence of values $c_1, c_2, \dots, c_N$ is chosen. It could be a sequence like

$0, 1, 0, 1, 0, 1, 0, 1, \ldots$, for example, which would be (if sufficiently long) very unlikely to occur if all c_n would be independently sampled from some distribution. This extends the possible experimental designs that we can handle to include for example randomized controlled trials in which the protocol is such that a certain prespecified number $N_{+|0}$ of subjects enters the control group, and a certain prespecified number $N_{+|1}$ enters the treatment group. If the values of c_n were chosen i.i.d. with a coin flip, then it would be very unlikely that this assignment satisfies the protocol. Considering C to be an exogenous input variable instead (with values that are not necessarily randomly assigned), allows us to test for the independence of outcome X and treatment C under a broader range of experimental protocols.

We define the *counts* as the number of observations with a given value $(x, c) \in \mathcal{X} \times \mathcal{C}$:

$$N_{x|c} := \sum_{n=1}^{N} \mathbb{1}_{(x,c)}(X_n, c_n).$$

We will take $H_0 : X \perp\!\!\!\perp C$ as the null hypothesis of a frequentist test for the independence of X and C. We parameterize the Markov kernel $K(X \mid C)$ in terms of parameters $(\theta_{x|c})_{x \in \mathcal{X}, c \in \mathcal{C}} := K(X = x \mid C = c)$ in a space

$$\mathcal{X}_\Theta := \{\theta \in \prod_{x \in \mathcal{X}, c \in \mathcal{C}} [0, 1] : \forall_{c \in \mathcal{C}} \sum_{x \in \mathcal{X}} \theta_{x|c} = 1\}.$$

With the summation convention, we can write the normalization condition as $\theta_{+|c} = 1$ for all $c \in \mathcal{C}$. The null hypothesis $H_0 : X \perp\!\!\!\perp C$, equivalent to (64), can be expressed in terms of the parameters as $H_0 : \theta \in \mathcal{X}_\Theta^0$, where we introduced the restricted parameter space

$$\mathcal{X}_\Theta^0 := \{\theta \in \mathcal{X}_\Theta : \forall_{x \in \mathcal{X}, c \in \mathcal{C}, c' \in \mathcal{C}} : \theta_{x|c} = \theta_{x|c'}\}.$$

As alternative hypothesis we will take $H_1 : X \not\perp\!\!\!\perp C$, the negation of the null hypothesis, i.e., $H_1 : \theta \in \mathcal{X}_\Theta^1$ with $\mathcal{X}_\Theta^1 := \mathcal{X}_\Theta \setminus \mathcal{X}_\Theta^0$.

We will again work out the likelihood ratio test. We first write down the "conditional" likelihood:

$$\theta \mapsto K\big((X_n)_{n=1}^N \mid (c_n)_{n=1}^N, \theta\big) = \prod_{n=1}^{N} \theta_{X_n|c_n} = \prod_{x \in \mathcal{X}} \prod_{c \in \mathcal{C}} \theta_{x|c}^{N_{x|c}}$$

where we used the counts as a sufficient statistic. Maximizing the likelihood with respect to the parameters, we obtain the maximum likelihood estimator

$$\hat{\theta}_{x|c} = \frac{N_{x|c}}{N_{+|c}},$$

i.e., the fractions of the different outcomes within each subgroup with $C = c$. Under H_0, we can write $\theta_{x|c} = \theta_{x|*}$ for some $\theta_{x|*} \in \mathcal{X}_\Theta^0$, and the likelihood simplifies:

$$\theta \in \mathcal{X}_\Theta^0 \implies \prod_{x \in \mathcal{X}} \prod_{c \in \mathcal{C}} \theta_{x|*}^{N_{x|c}} = \prod_{x \in \mathcal{X}} \theta_{x|*}^{N_{x|+}}.$$

The maximum likelihood estimator under H_0 is then

$$\hat{\theta}^0_{x|c} = \frac{N_{x|+}}{N}.$$

The likelihood ratio is obtained by dividing the likelihood for $\hat{\theta}$ by the likelihood for $\hat{\theta}^0$:

$$\frac{\sup_{\theta \in \mathcal{X}^0_\Theta \cup \mathcal{X}^1_\Theta} K((X_n)_{n=1}^N \mid (c_n)_{n=1}^N, \theta)}{\sup_{\theta \in \mathcal{X}^0_\Theta} K((X_n)_{n=1}^N \mid (c_n)_{n=1}^N, \theta)} = \prod_{x \in \mathcal{X}} \prod_{c \in \mathcal{C}} \left(\frac{\hat{\theta}_{x|c}}{\hat{\theta}^0_{x|c}}\right)^{N_{x|c}} = \prod_{x \in \mathcal{X}} \prod_{c \in \mathcal{C}} \left(\frac{N_{x|c}}{N_{+|c}} \frac{N}{N_{x|+}}\right)^{N_{x|c}}. \tag{65}$$

This is of *exactly* the same form as the likelihood ratio (59) for testing $X \perp\!\!\!\perp_{P(X,C)} C$ with two random variables X, C. The likelihood ratio test statistic is defined as 2 times the natural logarithm of this ratio:

$$G_N := 2 \log \frac{\sup_{\theta \in \mathcal{X}^0_\Theta \cup \mathcal{X}^1_\Theta} K((X_n)_{n=1}^N \mid (c_n)_{n=1}^N, \theta)}{\sup_{\theta \in \mathcal{X}^0_\Theta} K((X_n)_{n=1}^N \mid (c_n)_{n=1}^N, \theta)} = 2 \sum_{x \in \mathcal{X}} \sum_{c \in \mathcal{C}} N_{x|c} \log \frac{N_{x|c} N}{N_{x|+} N_{+|c}}$$

We (miraculously?!) arrived at the same test statistic as before. This time, we cannot make use of the general result on the asymptotic distribution under the null hypothesis of likelihood ratio tests. Indeed, that result pertains when dealing with a likelihood with only a finite number of parameters to be estimated, whereas here we have an asymptotically infinite number of "parameters" $c_1, c_2, \dots$. Therefore, we will resort to a more direct analysis of the case at hand. The end result will recover the previous results for random variables as a special case. Perhaps surprisingly, it turns out that under reasonable assumptions on the sequence $c_1, c_2, \dots$ we can apply the standard G test and ignore the non-random nature of the c_n's.

We will start with rewriting the test statistic. We introduce the space

$$\mathcal{X}_{\Theta_C} := \{\gamma \in \prod_{c \in \mathcal{C}} [0,1] : \sum_{c \in \mathcal{C}} \gamma_c = 1\}.$$

Consider now the function

$$\begin{aligned} g : \mathcal{X}_\Theta \times \mathcal{X}_{\Theta_C} : (\theta, \gamma) \mapsto &:= \sum_{c \in \mathcal{C}} \gamma_c \sum_{x \in \mathcal{X}} \theta_{x|c} \log \frac{\theta_{x|c}}{\sum_{c \in \mathcal{C}} \gamma_c \theta_{x|c}} \\ &= \sum_{c \in \mathcal{C}} \gamma_c \mathrm{KL}\Big(\theta_{X|c} \,\|\, \sum_{c \in \mathcal{C}} \gamma_c \theta_{X|c}\Big), \end{aligned} \tag{66}$$

where we introduced the notation $\theta_{X|c} := (\theta_{x|c})_{x \in \mathcal{X}} \in \mathbb{R}^{\mathcal{X}}$, and where the Kullback-Leibler divergence between two probability distributions $P, Q \in \mathcal{P}(\mathcal{X})$ is defined as:

$$\mathrm{KL}(P \,\|\, Q) := \sum_{x \in \mathcal{X}} P(X = x) \log \frac{P(X = x)}{Q(X = x)}$$

(and we identified a probability distribution on $\mathcal{X}$ with its probability mass function, encoded as a parameter vector in $\mathbb{R}^{\mathcal{X}}$). Note that

$$G_N = 2Ng(\hat{\theta}, \hat{\gamma})$$

with

$$\hat{\gamma}_c := \frac{N_{+|c}}{N}$$

for all $c \in \mathcal{C}$.[74] The components of $\hat{\gamma}$ are just the fractions of observations with a certain value of c. The asymptotic analysis will be conditional on the sequence of $\hat{\gamma}$'s, or equivalently, on the sequence of counts $N_{+|c}$.

Because the counts $(N_{x|c})_{x\in\mathcal{X}}$ have a multinomial distribution for each $c \in \mathcal{C}$, we can calculate that

$$\mathbb{E}\hat{\theta}_{x|c} = \frac{1}{N_{+|c}} \sum_{n=1}^{N} \mathbb{1}_c(c_n)\theta_{x|c} = \theta_{x|c}$$

and

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(\hat{\theta}_{x|c}, \hat{\theta}_{x'|c'}) &= \frac{1}{N_{+|c}N_{+|c'}} \sum_{n=1}^{N}\sum_{n'=1}^{N} \mathbb{C}\mathrm{ov}\big(\mathbb{1}_{(x,c)}(X_n, c_n)\mathbb{1}_{(x',c')}(X_{n'}, c_{n'})\big) \\
&= \frac{1}{N_{+|c}N_{+|c'}} \sum_{n=1}^{N} \mathbb{1}_c(c_n)\mathbb{1}_{c'}(c_n)\big(\delta_{xx'}\theta_{x|c} - \theta_{x|c}\theta_{x'|c}\big) \\
&= \frac{1}{N_{+|c}}\delta_{cc'}\theta_{x|c}(\delta_{xx'} - \theta_{x'|c}).
\end{aligned}$$

Assume that $N_{+|c} \to \infty$ for every $c \in \mathcal{C}$. Because

$$N_{x|c} = \sum_{\substack{n=1\\ c_n=c}}^{N} \mathbb{1}_x(X_n),$$

where the $\mathbb{1}_x(X_n)$ can be seen as i.i.d. vectors in $\mathbb{R}^{\mathcal{X}}$, we can apply the multivariate central limit theorem to conclude that

$$\sqrt{N_{+|c}}(\hat{\theta}_{x|c} - \theta_{x|c}) \rightsquigarrow \mathcal{N}(0, \Sigma) \tag{67}$$

with covariance matrix Σ with entries

$$(\Sigma)_{x|c,x'|c'} = \delta_{cc'}\theta_{x|c}(\delta_{xx'} - \theta_{x'|c}). \tag{68}$$

The central limit theorem thus provides the rate at which the ML estimate $\hat{\theta}$ converges to the true parameter θ. The likelihood ratio statistic G_N is a function of $\hat{\theta}$ and $\hat{\gamma}$. The asymptotic analysis of G_N can be obtained by performing a Taylor expansion of g around the true θ. This expansion can be done "uniformly" in $\hat{\gamma}$.

[74] We write $\hat{\gamma}$ rather than γ to indicate that it is an (N-dependent) function of the observed data, rather than a "true" fixed quantity.

Lemma 11.3.1. *Let $\theta \in \mathcal{X}^0_\Theta$ be positive, i.e., such that there exists a $\delta > 0$ with $\theta_{x|c} \geq \delta$ for all $x \in \mathcal{X}, c \in \mathcal{C}$. Let $\gamma \in \mathcal{X}_{\Theta_C}$. For $\hat{\theta}$ with $\|\hat{\theta} - \theta\|$ sufficiently small and $\hat{\theta} \xrightarrow{P} \theta$,*

$$g(\hat{\theta}, \gamma) = \frac{1}{2}(\hat{\theta} - \theta)^T \nabla^2_\theta g(\theta, \gamma)(\hat{\theta} - \theta) + o_P(\|\hat{\theta} - \theta\|^2), \tag{69}$$

with

$$\nabla^2_\theta g(\theta, \gamma) = \operatorname{diag}\left(\frac{1}{\theta_{X|*}}\right) \otimes \left(\operatorname{diag}(\sqrt{\gamma})\big(I_{\mathcal{C}} - \sqrt{\gamma}\sqrt{\gamma}^T\big)\operatorname{diag}(\sqrt{\gamma})\right)$$

where $\theta_{X|*} = (\theta_{x|c})_{x \in \mathcal{X}}$ for an arbitrary $c \in \mathcal{C}$.[75] *The remainder term $o_P(\|\hat{\theta} - \theta\|^2)$ can be chosen to be a function $f(\theta, \hat{\theta})$ that does not depend on γ, and for which*

$$\frac{f(\theta, \hat{\theta})}{\|\theta - \hat{\theta}\|^2} \xrightarrow{P} 0.$$

Proof. The second order Taylor expansion of g around the true θ is:

$$g(\theta + \epsilon, \gamma) = g(\theta, \gamma) + \epsilon^T \nabla_\theta g(\theta, \gamma) + \frac{1}{2}\epsilon^T \nabla^2_\theta g(\theta, \gamma)\epsilon + o(\|\epsilon\|^2),$$

where $\nabla_\theta g$ is the (partial) gradient of g with respect to θ, and $\nabla^2_\theta g$ is the Hessian of g with respect to θ, and the remainder term $o(\|\epsilon\|^2)$ can be taken of the form $M\|\epsilon\|^3$ if the third-order partial derivatives of g at (θ, γ) are bounded. For a random $\epsilon = \hat{\theta} - \theta$ we obtain

$$g(\hat{\theta}, \gamma) - g(\theta, \gamma) = (\hat{\theta} - \theta)^T \nabla_\theta g(\theta, \gamma) + \frac{1}{2}(\hat{\theta} - \theta)^T \nabla^2_\theta g(\theta, \gamma)(\hat{\theta} - \theta) + o_P(\|\hat{\theta} - \theta\|^2), \tag{70}$$

where the remainder term is now random, and converges in probability to 0 at rate $\|\hat{\theta} - \theta\|^2$. We will proceed by calculating the terms in the Taylor expansion.

The Kullback-Leibler divergence has the important property that $\mathrm{KL}(P \,\|\, Q) \geq 0$ and $\mathrm{KL}(P \,\|\, Q) = 0 \iff P = Q$ for all $P, Q \in \mathcal{P}(\mathcal{X})$. Together with the definition (66), this immediately implies that under H_0, $g(\theta, \gamma) = 0$ for all $\gamma \in \mathcal{X}_{\Theta_C}$.

The gradient $\nabla_\theta g$ of g w.r.t. θ has components:

$$\frac{\partial g}{\partial \theta_{x|c}} = \gamma_c \log \frac{\theta_{x|c}}{\sum_{c' \in \mathcal{C}} \gamma'_c \theta_{x|c'}}.$$

Under H_0, $\theta_{x|c} = \theta_{x|*}$ for all $x \in \mathcal{X}, c \in \mathcal{C}$, and it follows that the gradient vanishes for all $\gamma \in \mathcal{X}_{\Theta_C}$.

Next, the Hessian w.r.t. θ:

$$\frac{\partial^2 g}{\partial \theta_{x|c} \partial \theta_{x'|c'}} = \gamma_c \left(\frac{1}{\theta_{x|c}}\delta_{xx'}\delta_{cc'} - \frac{\gamma_{c'}}{\sum_{c'' \in \mathcal{C}} \gamma_{c''}\theta_{x|c''}}\delta_{xx'}\right).$$

[75] Here, we used the Kronecker product notation for matrices, and $\operatorname{diag}(v)$ is a diagonal matrix with the components of vector v on the diagonal.

Under H_0, $\theta_{x|c} = \theta_{x|*}$ for all $x \in \mathcal{X}, c \in \mathcal{C}$, this simplifies to

$$\frac{\partial^2 g}{\partial \theta_{x|c} \partial \theta_{x'|c'}} = \frac{1}{\theta_{x|*}} \gamma_c \left(\delta_{xx'} \delta_{cc'} - \gamma_{c'} \delta_{xx'} \right).$$

By using the Kronecker product notation, this can be written as stated in the lemma.

Finally, to obtain the remainder term, we calculate the third order partial derivatives:

$$\frac{\partial^3 g}{\partial \theta_{x|c} \partial \theta_{x'|c'} \partial \theta_{x''|c''}} = \gamma_c \delta_{x,x''} \delta_{x,x'} \left(-\frac{1}{\theta_{x|c}^2} \delta_{c,c''} \delta_{cc'} + \frac{\gamma_{c'} \gamma_{c''}}{(\sum_{c''' \in \mathcal{C}} \gamma_{c'''} \theta_{x|c'''})^2} \right).$$

These can be bounded uniformly in γ, using the assumption that all components of θ are bounded away from zero. $\square$

We are now ready to prove the following result on the asymptotic distribution of G_N under the null hypothesis, which is (surprisingly?) similar to Proposition 11.1.1.

Proposition 11.3.2. *Let $\theta \in \mathcal{X}_\Theta$ be positive, i.e., such that $\theta_{x|c} > 0$ for all $x \in \mathcal{X}, c \in \mathcal{C}$. Assume that $N_{+|c} \to \infty$ for all $c \in \mathcal{C}$. Under $H_0 : X \perp\!\!\!\perp_{K(X\,|\,C)} C$, the likelihood ratio test statistic (65) converges to a χ^2 distribution,*

$$G_N \rightsquigarrow \chi^2_\nu,$$

with $\nu = (|\mathcal{X}| - 1)(|\mathcal{C}| - 1)$ degrees of freedom.

Proof. We first note that $\hat{\theta} \overset{a.s.}{\to} \theta$ (for any $\theta \in \mathcal{X}_\Theta$) by applying the strong law of large numbers. Indeed, since the X_n are assumed to be conditionally i.i.d. given c_n, and

$$\mathbb{E}\big(\mathbb{1}_{(x,c)}(X_n, c_n)\big) = \theta_{x|c} \mathbb{1}_c(c_n)$$

for all $x \in \mathcal{X}, c \in \mathcal{C}$, and $N_{+|c} \to \infty$ for all $c \in \mathcal{C}$, we conclude that $N_{x|c}/N_{+|c} \overset{a.s.}{\to} \theta_{x|c}$ for all $x \in \mathcal{X}, c \in \mathcal{C}$ by the strong law of large numbers.

Since this implies convergence in probability $\hat{\theta} \overset{P}{\to} \theta$, Lemma 11.3.1 gives that under H_0,

$$g(\hat{\theta}, \hat{\gamma}) = \frac{1}{2} (\hat{\theta} - \theta)^T \nabla_\theta^2 g(\theta, \hat{\gamma}) (\hat{\theta} - \theta) + o_P(\|\hat{\theta} - \theta\|^2)$$

where

$$\nabla_\theta^2 g(\theta, \hat{\gamma}) = \operatorname{diag}\left(\frac{1}{\theta_{X|*}} \right) \otimes \left(\operatorname{diag}(\sqrt{\hat{\gamma}}) \big(I_{\mathcal{C}} - \sqrt{\hat{\gamma}} \sqrt{\hat{\gamma}}^T \big) \operatorname{diag}(\sqrt{\hat{\gamma}}) \right)$$

and the remainder term does not depend on $\hat{\gamma}$. Defining

$$S_N := \sqrt{N} \operatorname{diag}\left(\frac{1}{\sqrt{\theta_{X|*}}} \otimes \sqrt{\hat{\gamma}} \right) (\hat{\theta} - \theta)$$

where $\otimes$ denotes the Kronecker product of two vectors (in this case the all-ones vector $1_{\mathcal{X}} \in \mathbb{R}^{\mathcal{X}}$ and the vector $\sqrt{\hat{\gamma}} \in \mathbb{R}^{\mathcal{C}}$), and the orthogonal projection[76]

$$\hat{\Gamma} := I_{\mathcal{X}} \otimes \left(I_{\mathcal{C}} - \sqrt{\hat{\gamma}} \sqrt{\hat{\gamma}}^T \right),$$

[76] That is, $\hat{\Gamma}^T = \hat{\Gamma} = \hat{\Gamma}^2$.

we can write

$$N(\hat{\theta}-\theta)^T \nabla^2_\theta(\theta,\hat{\gamma})(\hat{\theta}-\theta) = S_N^T \hat{\Gamma} S_N = \|\hat{\Gamma} S_N\|^2.$$

Under H_0, $\theta_{x|c} = \theta_{x|*}$ for all $c \in \mathcal{C}$, and therefore (67) simplifies to:

$$\sqrt{N}\sqrt{\hat{\gamma}_c}(\hat{\theta}_{x|c} - \theta_{x|*}) \rightsquigarrow \mathcal{N}(0, \Sigma)$$

where the covariance matrix Σ from (68) simplifies to

$$(\Sigma)_{x|c,x'|c'} = \delta_{cc'}\theta_{x|*}(\delta_{xx'} - \theta_{x'|*}).$$

Hence, scaling with $\sqrt{\theta_{x|*}}$ gives

$$S_N = \sqrt{N}\sqrt{\hat{\gamma}_c}\frac{1}{\sqrt{\theta_{x|*}}}(\hat{\theta}_{x|c} - \theta_{x|*}) \rightsquigarrow \mathcal{N}(0, \tilde{\Sigma})$$

where the covariance matrix Σ from (68) simplifies to

$$(\tilde{\Sigma})_{x|c,x'|c'} = \delta_{cc'}\left(\delta_{xx'} - \sqrt{\theta_{x|*}}\sqrt{\theta_{x'|*}}\right),$$

which can be written using Kronecker product notation as

$$\tilde{\Sigma} = \left(I_{\mathcal{X}} - \sqrt{\theta_{X|*}}\sqrt{\theta^T_{X|*}}\right) \otimes I_{\mathcal{C}}.$$

Let $\hat{V} \in SO(\mathbb{R}^{\mathcal{C}})$ be rotations that map $\sqrt{\hat{\gamma}}$ to e_1. Then

$$I_{\mathcal{C}} - \sqrt{\hat{\gamma}}\sqrt{\hat{\gamma}}^T = P_{\sqrt{\hat{\gamma}}^\perp} = \hat{V}^T P_{e_1^\perp}\hat{V}$$

(where $P_{v^\perp}$ is the orthogonal projection on the subspace orthogonal to v), and therefore

$$\|\hat{\Gamma} S_N\|^2 = \|(I_{\mathcal{X}} \otimes \hat{V}^T P_{e_1^\perp}\hat{V})S_N\|^2 = \|(I_{\mathcal{X}} \otimes P_{e_1^\perp})(I_{\mathcal{X}} \otimes \hat{V})S_N\|^2.$$

We can apply Lemma 11.3.3 to conclude that $(I_{\mathcal{X}} \otimes \hat{V})S_N \rightsquigarrow \mathcal{N}(0, \tilde{\Sigma})$ as well (even though $\hat{V}$ is an N-dependent rotation). Then

$$(I_{\mathcal{X}} \otimes P_{e_1^\perp})(I_{\mathcal{X}} \otimes \hat{V})S_N \rightsquigarrow \mathcal{N}\left(0, P_{\sqrt{\theta_{X|*}}^\perp} \otimes P_{e_1^\perp}\right)$$

and hence

$$\|\hat{\Gamma} S_N\|^2 \rightsquigarrow \chi^2_\nu$$

with $\nu = (|\mathcal{X}| - 1)(|\mathcal{C}| - 1)$. With Slutsky's Lemma, also

$$G_N = 2Ng(\hat{\theta},\hat{\gamma}) = N(\hat{\theta}-\theta)^T \nabla^2_\theta(\theta,\hat{\gamma})(\hat{\theta}-\theta) + N o_P(\|\hat{\theta}-\theta\|^2) \rightsquigarrow \chi^2_\nu.$$

□

So, also the asymptotic distribution under the null hypothesis is the same as for the case of two random variables, even though we used a different parameterization, and we relaxed the assumption that the c_n's are i.i.d.: we only assumed that $N_{+|c} \to \infty$ for each c.[77] In particular, this result applies also to the case when C is a random variable. Hence, we have reobtained Proposition 11.1.1 as a special case.

Lemma 11.3.3. *Let Q be a rotationally symmetric probability measure on the standard Borel space $\mathbb{R}^k$ (i.e., $Q \circ U = Q$ for all $U \in SO(\mathbb{R}^k)$), and $P_1, P_2, \dots$ a sequence of probability measures on $\mathbb{R}^k$. Then*

$$P_n \rightsquigarrow Q \iff P_n \circ U_n \rightsquigarrow Q$$

for any sequence $U_1, U_2, \dots$ of rotations in $SO(\mathbb{R}^k)$.

Proof. We make use of the Lévy-Prokhorov metrization of the weak topology. For two probability distributions P, Q on $\mathbb{R}^k$ (with its Borel σ-algebra $\mathcal{B}_{\mathbb{R}^k}$), it is defined as

$$d(P,Q) := \inf\{\epsilon > 0 : \forall A \in \mathcal{B}_{\mathbb{R}^k} : P(A) \le Q(A^\epsilon) + \epsilon \ \wedge \ Q(A) \le P(A^\epsilon) + \epsilon\},$$

where $A^\epsilon := \cup_{a \in A} B_\epsilon(a)$ with $B_\epsilon(a) := \{v \in \mathbb{R}^k : \|v - a\| < \epsilon\}$. The Lévy-Prokhorov metric is invariant under rotations. Indeed, for $U \in SO(\mathbb{R}^k)$, we have that $UA \in \mathcal{B}_{\mathbb{R}^k} \iff A \in \mathcal{B}_{\mathbb{R}^k}$, and $(UA)^\epsilon = U(A^\epsilon)$ for $A \in \mathcal{B}_{\mathbb{R}^k}$ and $\epsilon > 0$, hence $d(P \circ U, Q \circ U) = d(P,Q)$ for all probability measures P, Q on $\mathbb{R}^k$. The rotation invariance of the Lévy-Prokhorov metric implies that if Q is rotationally symmetric (i.e., $Q = Q \circ U$ for all $U \in SO(\mathbb{R}^n)$), then

$$\begin{aligned}
P_n \rightsquigarrow Q &\iff d(P_n, Q) \to 0 \\
&\iff d(P_n \circ U_n, Q \circ U_n) \to 0 \\
&\iff d(P_n \circ U_n, Q) \to 0 \\
&\iff P_n \circ U_n \rightsquigarrow Q
\end{aligned}$$

□

Summarizing, we started from quite a different sampling scheme, and did not treat C as a random variable, yet we ended up with exactly the same likelihood ratio test for testing $X \perp\!\!\!\perp_{K(X|C)} C$ as we derived for testing the independence $X \perp\!\!\!\perp_{P(X,C)} C$ between two random variables.

What about the consistency of this test? To show consistency, it turns out that we need a stronger assumption than $N_{+|c} \to \infty$, but it will still be weaker than the i.i.d. assumption we made for the case that C is random.

Corollary 11.3.4. *Consider an infinite sequence of G tests performed on the first N samples of an infinitely large data set $(X_n, c_n)_{n=1}^\infty$, where one accepts $H_1 : X \not\perp\!\!\!\perp C$ if $G_N \ge \tau_N$, and otherwise accepts $H_0 : X \perp\!\!\!\perp C$, for some given sequence of thresholds τ_N.*

[77] If some of the $N_{+|c}$ stay finite asymptotically, then these c's can be ignored, and we still get asymptotically a chi-square distribution, but with less degrees of freedom.

Assume that $\theta \in \mathcal{X}_\Theta$ is positive, i.e., such that $\theta_{x|c} > 0$ for all $x \in \mathcal{X}, c \in \mathcal{C}$. Assume further that the fractions $N_{+|c}/N \to \infty$ are bounded away from zero asymptotically, i.e., there exists $\epsilon > 0$ such that for all $c \in \mathcal{C}$, $N_{+|c}/N \geq \epsilon$ for large N. Then this sequence of tests is asymptotically consistent if $\tau_N \to \infty$ but $\tau_N/N \to 0$.

Proof. In the proof of Proposition 11.3.2 we already saw that $\hat{\theta} \overset{a.s.}{\to} \theta$ for any $\theta \in \mathcal{X}_\Theta$.

Assume that H_1 holds, i.e., $\theta \in \mathcal{X}_\Theta^1$. Then $g(\theta, \gamma) = I(\gamma, \theta \circ \gamma) > 0$ for all $\gamma \in \mathcal{X}_{\Theta_C}$. $\mathcal{X}_\Theta^1$ is open in $\mathcal{X}_\Theta$, so $\bar{B}_\delta(\theta) \cap \mathcal{X}_\Theta = \{\tilde{\theta} \in \mathcal{X}_\Theta : \|\theta - \tilde{\theta}\| \leq \delta\} \subseteq \mathcal{X}_\Theta^1$ for δ small enough. Since $\hat{\theta} \overset{a.s.}{\to} \theta$, $\hat{\theta} \in \bar{B}_\delta(\theta) \cap \mathcal{X}_\Theta$ for large N a.s.. By assumption, for large N $\hat{\gamma} \in \{\tilde{\gamma} \in [\epsilon, 1]^{|\mathcal{C}|} : \tilde{\gamma}_+ = 1\}$, which is a closed subset of $\mathcal{X}_{\Theta_C}$. Since I is continuous, it attains a (positive) minimum value over the closed subset $(\bar{B}_\delta(\theta) \cap \mathcal{X}_\Theta) \times \{\tilde{\gamma} \in [\epsilon, 1]^{|\mathcal{C}|} : \tilde{\gamma}_+ = 1\} \subseteq \mathcal{X}_\Theta \times \mathcal{X}_{\Theta_C}$. Hence, $G_N/N = 2g(\hat{\theta}, \hat{\gamma})$ a.s. has a positive lower bound for large N. We can now reason analogously as in the proof of Corollary 11.1.3 to conclude that the probability of a Type II error vanishes asymptotically.

For an asymptotic estimate of the probability of a Type I error, we can make use of Proposition 11.3.2, which states that $G_N \rightsquigarrow \chi^2_\nu$ under H_0. This part of the proof is identical to the corresponding part of the proof of Corollary 11.1.3. □

The conditions in this corollary are sufficient, but not necessary. For example, not all the rates $N_{+|c}$ have to be lower bounded, it suffices if this is the case for a subset of $\mathcal{C}$ for which the distributions $K(X \mid C = c)$ differ. It also shows how consistency could fail: e.g., if the distributions $K(X \mid C = c)$ only differ on some subset of $\mathcal{C}$, but that subset is not observed sufficiently often asymptotically. This is in line with the intuition that when testing for the presence of a causal effect of C on X in a controlled setting (not necessarily randomized, i.e., as in Proposition 10.1.1), if nothing is known about how X might depend on C, it is best to gather sufficient data for each value that C can take.

11.4. The general categorical case

Finally, let us consider the most general case of testing a conditional independence involving transitional random variables (including "purely random" and "purely non-random" variables as special cases). Again, we will restrict ourselves to the case that all variables take values in finite spaces. We will formulate a general version of the G test.

Suppose we have three transitional random variables X, Y, Z and an input variable C, taking values in spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{C}$, respectively. Assume that all the spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{C}$ are finite. Suppose we have a kernel $K(X, Y, Z|C) : \mathcal{C} \dashrightarrow \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. We formulate a statistical test for testing the conditional independence

$$H_0 : X \underset{K(X,Y,Z|C)}{\perp\!\!\!\perp} Y \mid Z$$

against the alternative

$$H_1 : X \underset{K(X,Y,Z|C)}{\not\!\perp\!\!\!\perp} Y \mid Z.$$

The null hypothesis is equivalent, by definition, to the existence of a kernel $K(X|Z)$ such that

$$K(X, Y, Z \mid C) = K(X \mid Z) \otimes K(Y, Z \mid C).$$

It will turn out to be helpful to consider the equivalent hypotheses

$$H_0 : X \underset{K(X,Y,Z|C)}{\perp\!\!\!\perp} Y, C \mid Z$$

against the alternative

$$H_1 : X \underset{K(X,Y,Z|C)}{\not\!\perp\!\!\!\perp} Y, C \mid Z$$

instead.

Suppose we obtain data $(X_n, Y_n, Z_n, c_n)_{n=1}^N$ such that the (X_n, Y_n, Z_n) are conditionally independent and identically distributed given c_n for all $n = 1, \ldots, N$. In other words, we assume the data is sampled from the following Markov kernel:

$$K((X_n, Y_n, Z_n)_{n=1}^N \mid (c_n)_{n=1}^N) = \bigotimes_{n=1}^N K(X_n, Y_n, Z_n \mid C_n = c_n),$$

where each $K(X_n, Y_n, Z_n \mid C_n)$ is a copy of the ("true" but unknown kernel) $K(X, Y, Z \mid C)$. We parameterize this kernel $K(X, Y, Z \mid C)$ as

$$K(X = x, Y = y, Z = z \mid C = c) = \theta_{xyz|c}$$

with θ in

$$\mathcal{X}_\Theta = \{\theta \in \prod_{(x,y,z,c) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \times \mathcal{C}} : \theta_{+++|c} = 1 \ \forall c \in \mathcal{C}\}.$$

We will not work out the details, as these are analogous to what we have seen before, but will directly formulate the likelihood ratio test statistic:

$$G_N = 2 \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} \sum_{c \in \mathcal{C}} N_{xyzc} \log \frac{N_{xyzc} N_{++z+}}{N_{x+z+} N_{+yzc}}. \tag{71}$$

With the correspondence $(X, (Y, C), Z) \leftrightarrow (X, Y, Z)$, this likelihood ratio statistic is seen to be identical to (62), the one for the case of three random variables. Note that this likelihood ratio treats C and Y on equal footing. This, again, suggests that for the asymptotic analysis we will obtain a similar result as that for the conditional G test with purely random variables, albeit under milder assumptions on the sampling scheme for C. We will not work out the details here, but directly formulate the result.

Proposition 11.4.1. *Let $\theta \in \mathcal{X}_\Theta$ be positive, i.e., such that $\theta_{xyz|c} > 0$ for all $x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}, c \in \mathcal{C}$. Assume that $N_{+++|c} \to \infty$ for all $c \in \mathcal{C}$. Under $H_0 : X \perp\!\!\!\perp_{K(X,Y,Z \mid C)} Y, C \mid Z$, the likelihood ratio test statistic (71) converges to a χ^2 distribution,*

$$G_N \rightsquigarrow \chi^2_\nu,$$

with $\nu = |\mathcal{Z}| \, (|\mathcal{X}| - 1)(|\mathcal{Y}||\mathcal{C}| - 1)$ degrees of freedom.

Proof. Note that the likelihood ratio test statistic (71) is a sum over $z \in \mathcal{Z}$ of a likelihood ratio test statistic of the form (65) (where (Y, C) in the former corresponds with C in the latter). □

We also obtain a similar result as before on the asymptotic consistency.

Corollary 11.4.2. *Consider an infinite sequence of G tests performed on the first N samples of an infinitely large data set $(X_n, Y_n, Z_n, c_n)_{n=1}^{\infty}$, where one decides*

$$\begin{cases} H_0 : X \perp\!\!\!\perp_{K(X,Y,Z\,|\,C)} Y, C \,|\, Z & \text{if } G_N < \tau_N, \\ H_1 : X \not\perp\!\!\!\perp_{K(X,Y,Z\,|\,C)} Y, C \,|\, Z & \text{if } G_N \geq \tau_N, \end{cases}$$

for some given sequence of thresholds τ_N. Assume that $\theta \in \mathcal{X}_\Theta$ is positive, i.e., such that $\theta_{xyz|c} > 0$ for all $x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}, c \in \mathcal{C}$. Assume further that the fractions $N_{+++|c}/N \to \infty$ are bounded away from zero asymptotically, i.e., there exists $\epsilon > 0$ such that for all $c \in \mathcal{C}$, $N_{+++|c}/N > \epsilon$ for large N. Then this sequence of tests is asymptotically consistent if $\tau_N \to \infty$ but $\tau_N/N \to 0$.

12. The Fast Causal Inference Algorithm

In this final chapter, we present an extension of the Fast Causal Inference (FCI) algorithm. The FCI algorithm is one of the highlights in the field of constraint-based causal discovery. It was originally designed for purely observational (non-experimental) data and relied on the assumption of acyclicity, but allowed for latent variables (either marginalized over, or conditioned upon in case of selection bias) [SMR95, SMR99]. Later, the algorithm was augmented with additional 'orientation rules' and this augmented FCI algorithm was shown to be complete in a certain sense [Zha08].[78] While the FCI algorithm was originally designed for the acyclic setting, it was recently discovered that it also works in case cycles are present (more specifically, for σ-faithful simple SCMs) [MC20]. While that work made a simplification by assuming no selection bias, we prove here that the FCI algorithm is sound even when the data is generated according to a conditional Markov kernel induced by a σ-faithful simple SCM, or in other words, that it can cope with the possible presence of cycles *and* selection bias. In other recent work, the FCI algorithm has been extended to incorporate prior knowledge regarding exogeneity and unconfoundedness of context variables (of the same type that we used for modeling the treatment variable in randomized controlled trials) [MMC20]. Such an extension with exogenous input nodes allows to generalize the idea of causal discovery in a randomized controlled trial to multiple treatment and outcome variables. We also incorporate that extension here, and provide a 'unified' extended FCI algorithm that allows for cycles (under certain assumptions), selection bias and exogenous input variables.

12.1. Modeling selection bias

We model selection bias as follows.

Notation 12.1.1. *We will assume the existence of a simple SCM $M = (J, V^+, W, \mathcal{X}, P, f)$, with endogenous variables $V^+ = V \dot\cup S$.*[79] *We assume that only the endogenous variables in V are observed, as well as the exogenous input variables J. The latent variables in S have the role of* ***latent selection variables****. In other words, we will assume that the data is distributed according to the marginal Markov kernel of M on V after conditioning on S:*

$$P_M(X_V \mid X_S \in \xi_S, \mathrm{do}(X_J))$$

for some measurable set $\xi_S \subseteq \mathcal{X}_S$.[80]

[78] The augmented version of [Zha08] is often referred to simply as 'the FCI algorithm', and we will do so here as well.

[79] Alternatively, one could assume endogenous variables $V^+ = V \dot\cup S \dot\cup L$ where S are used as selection variables. We could then start by marginalizing out the variables in L to arrive at the setting that is our starting point here.

[80] Note that we did not include the exogenous random variables here; we are treating them as if they were unobserved. If some (or all) exogenous random variables are observed, then we have three options: (i) include them, but ignore the additional background knowledge that they are independent (that

This models a process that *filters* the data according to the value of X_S (like in a rejection sampler).[81] One practical example of such a filtering process leading to selection bias is the following.

Example 12.1.2. *After the exam, the teacher hands out evaluation forms to the students and asks them to evaluate the course. When analyzing this evaluation, the teacher needs to be aware of possible selection bias. For example, if the students that were most dissatisfied with the course already dropped out earlier on and never filled in the evaluation form, the observed evaluations may not be representative of the opinions of all the students that started the course. Here, the selection variable may be the binary indicator 'student filled in the evaluation form'.*

Our goal in the rest of this chapter will be to deduce as much as possible about the graph $G_{V^+|J}(M)$ from the Markov kernel $P_M(X_V \mid X_S \in \xi_S, \mathrm{do}(X_J))$. FCI is an example of a *constraint-based* causal discovery algorithm, which means that it only utilizes the conditional independence information of this Markov kernel.

Proposition 12.1.3. *Given a Markov kernel $K(X_V \mid X_J)$. Suppose there exists a simple SCM $M = (J, V^+, W, \mathcal{X}, P, f)$, with endogenous variables $V^+ = V \dot{\cup} S$, such that*

$$K(X_V \mid X_J) = P_M(X_V \mid X_S \in \xi_S, \mathrm{do}(X_J))$$

for some measurable set $\xi_S \subseteq \mathcal{X}_S$. Then there exists a simple SCM $\tilde{M} = \left(J, \tilde{V}^+, W, \mathcal{X}, P, \tilde{f}\right)$, with endogenous variables $\tilde{V}^+ = V \dot{\cup} \{\tilde{s}\}$, such that

$$P_M(X_V \mid X_S \in \xi_S, \mathrm{do}(X_J)) = P_{\tilde{M}}(X_V \mid X_{\tilde{s}} = 1, \mathrm{do}(X_J))$$

and $\mathrm{Anc}_{G(M)}(S) = \mathrm{Anc}_{G(\tilde{M})}(\tilde{s})$.

Proof. Define $\hat{M} = \left(J, \hat{V}^+, W, \mathcal{X}, P, \hat{f}\right)$, with endogenous variables $\hat{V}^+ = V \dot{\cup} S \dot{\cup} \{\tilde{s}\}$, and $\hat{f}_V := f_V$, $\hat{f}_S := f_S$, $\hat{f}_{\tilde{s}}(x) := \mathbb{1}_{\xi_S}(x_S)$. Now take $\tilde{M} = \hat{M}^{\setminus S}$. $\square$

12.2. Inducing walks

The key notion in this chapter is that of *σ-inducing walks.* We define σ-inducing walk as a generalization to CDMGs of the notion of inducing path [VP90] in DAGs.

Definition 12.2.1. *Let G be a CDMG with input nodes J, output nodes $V^+ = V \dot{\cup} S$ and let $i, j \in V \dot{\cup} J$ be distinct nodes. A walk π in G between i and j is called **σ-inducing given** S if each collider on π is in $\mathrm{Anc}_G(\{i,j\} \cup S)$, and each non-endpoint non-collider*

is, treating them as if they were endogenous), (ii) make endogenous copies and include those; (iii) include them and make use of the additional background knowledge that they are independent. Here, we chose for the second option.

[81] It is helpful to think about such a filtering step as an intervention on the population level, but may lead to confusion when interpreted as an intervention on the individual level.

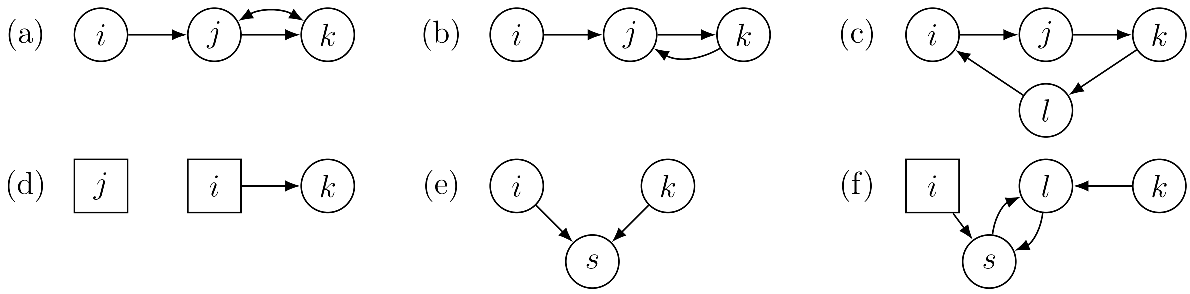


Figure 30: Examples of non-trivial σ-inducing paths in CDMGs. (a) The path $i \to j \leftrightarrow k$ is a σ-inducing path between i and k. (b) The path $i \to j \to k$ is a σ-inducing path between i and k. (c) The path $i \to j \to k$ is a σ-inducing path between i and k. The nodes i and k cannot be σ-separated by any subset not containing i, k (indeed, $i \not\perp^{\sigma} k$ and $i \not\perp^{\sigma} k \mid j$ in all three graphs, and additionally $i \not\perp^{\sigma} k \mid l$ and $i \not\perp^{\sigma} k \mid \{j, l\}$ in the graph in (c)). (d) The path $i \to k$ is σ-inducing, while there is no σ-inducing path between i and j, nor between j and k. (e) The path $i \to s \leftarrow k$ is σ-inducing given $\{s\}$. (f) The path $i \to s \to l \leftarrow k$ is σ-inducing given $\{s\}$.

on π is unblockable. If it is a path, it is called a ***σ-inducing path given*** *S between i and j.*[82]

If two nodes are adjacent in G, any edge connecting the two is a σ-inducing walk (path) given S between them, for any S. Figure 30 shows some simple nontrivial examples of σ-inducing paths.

Lemma 12.2.2. *Let G be a CDMG with input nodes J, output nodes $V^+ = V \dot\cup S$ and let $i, j \in V \dot\cup J$ be distinct nodes. If $i \in \mathrm{Sc}_G(j)$ then there exists a σ-inducing path given S in G between i and j.*

Proof. There exists a directed path in G from i to j that is entirely contained in $\mathrm{Sc}_G(j)$, and therefore all its non-endpoint nodes are unblockable non-colliders. $\square$

The notion of σ-inducing walk has the following important properties.

Proposition 12.2.3. *Let G be a CDMG with input nodes J, output nodes $V^+ = V \dot\cup S$ and let $i \in V$ (but $i \notin J$) and $j \in V \cup J$ be distinct nodes. Then the following are equivalent:*

(i) There is a σ-inducing path given S in G between i and j;

(ii) There is a σ-inducing walk given S in G between i and j;

(iii) $i \not\perp^{\sigma}_{G} j \mid S \cup Z$ for all $Z \subseteq (V \cup J) \setminus \{i, j\}$;

[82] In most of the literature, the graph is assumed to be acyclic, and then the notion is referred to simply as "inducing".

(iv) $i \not\perp^{\sigma}_{G} j \mid S \cup Z$ *for* $Z = (\mathrm{Anc}_G(\{i,j\} \cup S) \cup J) \setminus \{i,j\}$.

Proof. The proof is similar to that of Theorem 4.2 in [RS02].

(i) $\Longrightarrow$ (ii) is trivial.

(ii) $\Longrightarrow$ (iii): Assume the existence of a σ-inducing walk given S between i and j in G. Let $Z \subseteq (V \cup J) \setminus \{i,j\}$. Consider all walks in G between i and j with the property that all colliders on it are in $\mathrm{Anc}_G(\{i,j\} \cup S \cup Z)$, and each non-endpoint non-collider on it is not in $S \cup Z$ or is unblockable. Such walks exist, since the σ-inducing walk is one. Let μ be such a walk with a minimal number of colliders. We show that all colliders on μ must be in $\mathrm{Anc}_G(S \cup Z)$. Suppose on the contrary the existence of a collider k on μ that is not ancestor of $S \cup Z$. It is either ancestor of i or of j, by assumption. If $j \in J$, it cannot be ancestor of j, and hence must be ancestor of i. Otherwise, we can assume it to be ancestor of i without loss of generality. Then there is a directed path π from k to i in G that does not pass through any node of $S \cup Z$. Then the subwalk of μ between k and j can be concatenated with the directed path π into a walk between i and j that has the property, but has fewer colliders than μ: a contradiction. Therefore, μ is σ-open given $S \cup Z$. Hence, i and j are σ-connected given $S \cup Z$.

(iii) $\Longrightarrow$ (iv) is trivial.

(iv) $\Longrightarrow$ (i): Suppose that i and j are σ-connected given $Z = (\mathrm{Anc}_G(\{i,j\} \cup S) \cup J) \setminus \{i,j\}$. Let π be a path between i and $\{j\} \cup J$ that is σ-open given Z. The end nodes of π must be i and j, because $J \setminus \{i,j\} \subseteq Z$. We show that π must be a σ-inducing path given S. First, all colliders on π are in $\mathrm{Anc}_G(Z)$, but not in J, and hence in $\mathrm{Anc}_G(\{i,j\} \cup S)$. Second, let k be any non-endpoint non-collider on π. Then there must be a directed subpath of π starting at k that ends either at the first collider on π next to k or at an end node of π, and hence k must be in Z. Since π is σ-open given Z, k must be unblockable. Hence, all non-endpoint non-colliders on π must be unblockable. $\square$

In words: there is a σ-inducing path between two nodes in a CDMG (provided they are not both input nodes) if and only if the two nodes cannot be σ-separated by any subset of the other nodes.

Remark 12.2.4. *Two input nodes cannot be σ-separated by some subset of other nodes. Indeed, the trivial path j is σ-open as long as we don't condition on j itself. On the other hand, there may, or may not, be a σ-inducing path given S between two input nodes. For example, in the CDMG $j_1 \rightarrow i \leftarrow j_2$ with $j_1, j_2 \in J$ the path $j_1 \rightarrow i \leftarrow j_2$ is the only path between j_1 and j_2, and it is is σ-inducing given $\{i\}$ but not σ-inducing given $\emptyset$. This is why Proposition 12.2.3 does not consider the case $i, j \in J$.*

The orientations of the outermost edges on a σ-inducing path contain important information about ancestral relations.

Lemma 12.2.5. *Let G be a CDMG with input nodes J, output nodes $V^+ = V \dot\cup S$ and let $i, j \in V \dot\cup J$ be distinct. If there exists a σ-inducing path given S between i and j in G, and all σ-inducing paths given S in G between i and j are out of j, then $j \in \mathrm{Anc}_G(\{i\} \cup S)$.*

Proof. Let μ be a σ-inducing path given S between i and j in G. It must be of the form $i \cdots l \leftarrow j$ (with possibly $l = i$). First we show that l cannot be in $\mathrm{Sc}_G(j)$. If $l \in \mathrm{Sc}_G(j)$, then let π be a directed path in G from l to j that is entirely contained in $\mathrm{Sc}_G(j)$. Let m be the node on μ closest to i that is also on π (possibly $m = l$). The subpath of π between j and m can be concatenated with the subpath of μ between m and i into a walk between j and i. This must be a σ-inducing path given S between i and j that is into j by construction: contradiction. Hence l cannot be in $\mathrm{Sc}_G(j)$.

If μ is a directed path all the way to i, then clearly, $j \in \mathrm{Anc}_G(\{i\} \cup S)$. Otherwise, it must contain a collider. Let k be the collider on μ closest to j. k must be ancestor of i or j or S. In the first and third cases, clearly $j \in \mathrm{Anc}_G(\{i\} \cup S)$. In the second case, all nodes on the subpath of μ between j and k must be in $\mathrm{Sc}_G(j)$, a contradiction. □

Lemma 12.2.6. *Let G be a CDMG with input nodes J, output nodes $V^+ = V \dot\cup S$ and let $i, j \in V \dot\cup J$ be distinct. If there exists a σ-inducing path given S between i and j in G into j, and $i \notin \mathrm{Anc}_G(\{j\} \cup S)$, then there exists a σ-inducing path given S between i and j in G that is both into i and into j.*

Proof. Let μ be a σ-inducing path given S between i and j in G into j. This rules out $j \in J$. If μ is into i, we are done. Therefore, suppose it is of the form $i \rightarrow \cdots \ast\!\!\rightarrow j$. It cannot be a directed path, since $i \notin \mathrm{Anc}_G(\{j\} \cup S)$. Therefore, there must be a collider k on μ such that μ is of the form $i \rightarrow \ldots \rightarrow k \leftarrow\!\!\ast \cdots \ast\!\!\rightarrow j$ (with the subpath between i and k directed). Then $k \in \mathrm{Anc}_G(\{i\})$ (sic!), and hence all nodes on μ between i and k must be in $\mathrm{Sc}_G(i)$. Let π be a directed path in G from k to i that is entirely contained in $\mathrm{Sc}_G(i)$. Let l be the node on μ closest to j that is also on π (possibly $l = k$). Then $l \neq j$, because otherwise $j \in \mathrm{Sc}_G(i)$, contradicting $i \notin \mathrm{Anc}_G(\{j\} \cup S)$. The non-trivial subpath of π between i and l can be concatenated with the non-trivial subpath of μ between l and j into a walk between i and j. This must be a σ-inducing path given S between i and j that is both into i and into j. □

Lemma 12.2.7. *Let G be a CDMG with input nodes J, output nodes $V^+ = V \dot\cup S$ and let $i, j \in V \dot\cup J$ be distinct. If there is a σ-inducing path in G given S between i and j that is into j, and $k \in \mathrm{Sc}_G(j)$, then there is a σ-inducing path in G given S between i and k that is into k.*

Proof. The σ-inducing path given S in G between i into j can be extended by a directed path within $\mathrm{Sc}_G(k)$ to a σ-inducing walk given S in G between i and k that is into k. □

Lemma 12.2.8. *Let G be a CDMG with input nodes J, output nodes $V^+ = V \dot\cup S$. If there is a σ-inducing path in G given S between $j \in J$ and $v \in V$, then $j \in \mathrm{Anc}_G(\{v\} \cup S)$.*

Proof. The σ-inducing path in G must start out of j by definition. If it is a directed path from j all the way to v, then clearly $j \in \mathrm{Anc}_G(v) \subseteq \mathrm{Anc}_G(\{v\} \cup S)$. Otherwise, it will contain a directed path from j to a collider. That collider must be in $\mathrm{Anc}_G(\{j, v\} \cup S)$, but since it cannot be in $\mathrm{Anc}_G(j)$, it has to be in $\mathrm{Anc}_G(\{v\} \cup S)$. □

12.3. Partial Ancestral Graphs

It is often convenient when performing causal reasoning to be able to represent a set of CDMGs in a compact way. For this purpose, *partial ancestral graphs (PAGs)* have been introduced [SMR99, Zha06]. In order to deal with possible cycles (in simple SCMs), selection bias and exogenous input nodes, we extend the definition to σ-PAGs.

σ-PAGs have two node types, input nodes and output nodes (just like CDMGs). They also have multiple edge types. In addition to the three edge types for CDMGs ($\to$, $\leftarrow$, $\leftrightarrow$), there is an undirected edge ($-$) and there are five edge types involving a circle edge mark: $\leftarrow\!\circ$, $\circ\!-\!\circ$, $\circ\!\!\to$, $-\!\circ$, $\circ\!-$. Each edge $i \ast\!-\!\ast\, j$ has two *edge marks*, one at each node, with each edge mark either a *tail*, *arrowhead* or *circle*. For example, the directed edge $i \to j$ has a tail at i and an arrowhead at j, while the bi-circle edge $i \circ\!-\!\circ\, j$ has two circle edge marks. All 9 possible combinations of edge marks can occur on an edge in a σ-PAG. We will make use of the "$*$" symbol to denote any of the three edge marks. So the notation $i \ast\!-\!\ast\, j$ can stand for all 9 possible edge types between i and j, whereas $i \leftarrow\!\ast\, j$ is shorthand for three possible edge types, as are $i -\!\ast\, j$ and $i \circ\!-\!\ast\, j$. Edges of the form $i \leftarrow\!\ast\, j$ and $j \ast\!\!\to i$ are called *into* i. Edges of the form $i -\!\ast\, j$ and $j \ast\!- i$ are called *out of* i.

In order to define σ-PAGs, we extend the definitions of (directed) walks, (directed) paths and colliders to cover these new edge types.

Definition 12.3.1. *Let $H = (J, V, E)$ be a mixed graph with input nodes J, output nodes V and edges E of the types $\{\to, \leftarrow, \leftarrow\!\circ, \leftrightarrow, \circ\!-\!\circ, \circ\!\!\to, -\!\circ, \circ\!-, -\}$.*[83] *Let $v, w \in V \cup J$.*

1. *If there is an edge $v \ast\!-\!\ast\, w$ between v and w (of any type), we call v and w **adjacent** in H.*
2. *Define $\mathrm{Adj}_H(v)$ to be the nodes adjacent to v in H.*
3. *A triple of distinct nodes (a, b, c) in H form a **triangle** if each pair of nodes in the triple is adjacent in H.*
4. *A triple of distinct nodes (a, b, c) in H is called **unshielded** if b is adjacent to both a and c in H, but a is not adjacent to c in H.*
5. *A **walk between** v **and** w in H is a finite alternating sequence of nodes and edges*

$$v = v_0, a_0, v_1, \ldots v_{n-1}, a_{n-1}, v_n = w$$

in H for some $n \geq 0$, i.e. such that for every $k = 0, \ldots, n-1$ we have that $v_k, v_{k+1} \in V \cup J$ and $a_k = v_k \ast\!-\!\ast\, v_{k+1} \in E$, and with end nodes $v_0 = v$ and $v_n = w$.

6. *A walk is called a **path** if no node occurs more than once on the walk.*
7. *A path $v_0 \ast\!-\!\ast \ldots \ast\!-\!\ast\, v_n$ in H is called **uncovered** if every subsequent triple (v_{k-1}, v_k, v_{k+1}) for $1 < k < n$ is unshielded.*

[83] Formally, we no longer introduce separate sets to represent the edges of each type, but merge them into the single set E.

We also introduce a new type of path that is specific for mixed graphs that allow circle edge marks.

Definition 12.3.2. *Let $H = (J, V, E)$ be a mixed graph with input nodes J, output nodes V and edges E of the types $\{\to, \leftarrow, \leftarrow\!\!\circ, \leftrightarrow, \circ\!\!-\!\!\circ, \circ\!\!\to, -\!\!\circ, \circ\!\!-, -\}$.*

1. *A path of the form $v_0 \circ\!\!-\!\!\circ \dots \circ\!\!-\!\!\circ v_n$ (with each edge of the form $v_i \circ\!\!-\!\!\circ v_{i+1}$) is called a **circle path**.*

By interpreting a circle edge mark as an edge mark that could become a tail or an arrowhead, we obtain also the following definitions.

Definition 12.3.3. *Let $H = (J, V, E)$ be a mixed graph with input nodes J, output nodes V and edges E of the types $\{\to, \leftarrow, \leftarrow\!\!\circ, \leftrightarrow, \circ\!\!-\!\!\circ, \circ\!\!\to, -\!\!\circ, \circ\!\!-, -\}$.*

1. *A triple of consecutive nodes $v_{k-1} \ast\!\!-\!\!\ast v_k \ast\!\!-\!\!\ast v_{k+1}$ on a walk in H is called **definite collider** if it is of the form $v_{k-1} \ast\!\!\to v_k \leftarrow\!\!\ast v_{k+1}$.*

2. *A triple of consecutive nodes $v_{k-1} \ast\!\!-\!\!\ast v_k \ast\!\!-\!\!\ast v_{k+1}$ on a walk in H is called a **definite non-collider** if it is of the form $v_{k-1} \ast\!\!-\!\!\ast v_k -\!\!\ast v_{k+1}$ or $v_{k-1} \ast\!\!- v_k \ast\!\!-\!\!\ast v_{k+1}$. Furthermore, we also refer to the end nodes v_0 and v_n of a walk between v_0 and v_n in H as **definite non-colliders**.*

3. *A path between v and w in H is called a **definite collider path** if every non-endpoint node on the path is a definite collider on the path.*

4. *A walk between $v, w \in V \cup J$ (with $v \neq w$) in H is called **definitely inducing** if every non-endpoint node is a definite collider in $\mathrm{Anc}_H(\{v, w\})$.*

5. *A path $v_0 \ast\!\!-\!\!\ast \dots \ast\!\!-\!\!\ast v_n$ in H is called a **possibly directed path from** v_0 **to** v_n if for each $i = 1, \dots, n$, the edge $v_{i-1} \ast\!\!-\!\!\ast v_i$ is not into v_{i-1} and is not out of v_i (i.e., each edge must be of the form $v_{i-1} \circ\!\!-\!\!\circ v_i$, $v_{i-1} \circ\!\!\to v_i$, $v_{i-1} -\!\!\circ v_i$ or $v_{i-1} \to v_i$).*

We can now define:[84]

Definition 12.3.4. *A mixed graph $H = (J, V, E)$ with input nodes J, output nodes V and edges E of the types $\{\to, \leftarrow, \leftarrow\!\!\circ, \leftrightarrow, \circ\!\!-\!\!\circ, \circ\!\!\to, -\!\!\circ, \circ\!\!-, -\}$ is called a **partial σ-ancestral graph (σ-PAG)** if all of the following conditions hold:*

1. *Between any two distinct nodes there is at most one edge, and there are no edges between a node and itself;*

2. *No input node is adjacent to any other input node;*

[84] We have incorporated two extensions of the usual definition of PAG [Zha06]: we allow for input nodes, and we have weakened the condition of being ancestral to σ-ancestral. A mixed graph is called *ancestral* if it has no directed cycles, no almost directed cycles, and no triples of the form $i \ast\!\!\to j - k$.

3. *The graph contains no directed cycles, no almost directed cycles, and no unshielded triple of the form $i \ast\!\!\rightarrow j - k$ ("σ-ancestral");*

4. *There is no definitely inducing path between any two distinct non-adjacent nodes ("maximal");*

σ-PAGs are used to represent a set of CDMGs as follows.

Definition 12.3.5. *Let $H = (J, V, E)$ be a mixed graph with input nodes J, output nodes V and edges E of the types $\{\rightarrow, \leftarrow, \leftarrow\!\circ, \leftrightarrow, \circ\!\!-\!\!\circ, \circ\!\!\rightarrow, -\!\!\circ, \circ\!\!-, -\}$. Let G be a CDMG with input nodes J and output nodes $V^+ = V \dot\cup S$. We say that H **represents** G **given** S if all of the following hold:*

1. *Between any two distinct nodes in H there is at most one edge, and there are no edges between a node and itself;*

2. *Two distinct nodes $i, j \in V \cup J$ are adjacent in H if and only if $\{i, j\} \not\subseteq J$ and there is a σ-inducing path between i and j given S in G;*

3. *If $i \leftarrow\!\!\ast\, j$ in H then $i \notin \mathrm{Anc}_G(\{j\} \cup S)$;*

4. *If $i -\!\!\ast\, j$ in H then $i \in \mathrm{Anc}_G(\{j\} \cup S)$;*

Hence, adjacencies represent σ-inducing paths, arrowheads represent non-ancestorship (of the adjacent node or S), and tails represent ancestorship (of the adjacent node or S). Some examples are given in Figure 31. Note in particular that a directed edge in a σ-PAG does not necessarily imply a direct causal relationship (for example, the edge $i \rightarrow k$ in Figure 31(a)). Figure 32 provides an example of a mixed graph that satisfies all conditions of a σ-PAG except the maximality.

The following elementary properties shows that certain important properties of the CDMG are encoded in a mixed graph that represents it.

Lemma 12.3.6. *Let $H = (J, V, E)$ be a mixed graph with input nodes J, output nodes V and edges E of the types $\{\rightarrow, \leftarrow, \leftarrow\!\circ, \leftrightarrow, \circ\!\!-\!\!\circ, \circ\!\!\rightarrow, -\!\!\circ, \circ\!\!-, -\}$. Let G be a CDMG with input nodes J and output nodes $V^+ = V \dot\cup S$. Suppose that H represents G given S. Then for any two nodes i, j in H:*

(i) $i \in \mathrm{Anc}_H(j)$ implies $i \in \mathrm{Anc}_G(\{j\} \cup S)$.

(ii) If $i \ast\!\!\rightarrow j$ in H, then there exists a σ-inducing path given S in G between i and j that is into j.

(iii) If $i \leftrightarrow j$ in H, then there exists a σ-inducing path given S in G between i and j that is both into i and into j.

Proof. (i) Suppose H contains a directed path $i = v_0 \rightarrow \ldots \rightarrow v_n = j$. Note that for all $k = 0, \ldots, n-1$, $v_k \notin \mathrm{Anc}_G(S)$ implies $v_k \in \mathrm{Anc}_G(v_{k+1})$. By induction, then, $v_0 \notin \mathrm{Anc}_G(S)$ implies $v_0 \in \mathrm{Anc}_G(v_n)$.

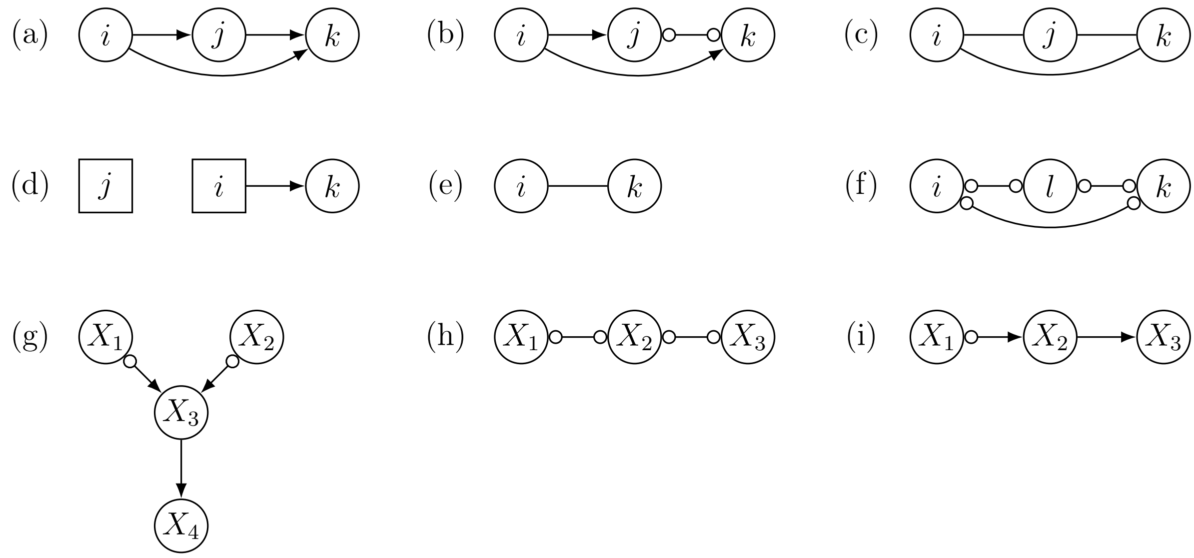


Figure 31: Various example σ-PAGs, representing respectively the CDMGs: (a) in Figure 30(a); (b) in Figures 30(a–b); (c) in Figures 30(a–b); (d) in Figure 30(d); (e) in Figure 30(e); (f) in Figure 30(f); (g) of all Y-structures in Figure 29; (h) of all LCD structures in Figure 27; (i) of all LCD structures in Figure 27.

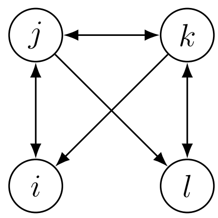


Figure 32: This mixed graph is not a valid σ-PAG, because it is not maximal: it has a definitely inducing path $i \leftrightarrow j \leftrightarrow k \leftrightarrow l$ while i and l are non-adjacent.

(ii) There exists a σ-inducing path given S between i and j in G because i and j are adjacent in H and H represents G given S. If all σ-inducing paths given S between i and j in G were out of j, then by Lemma 12.2.5, $j \in \mathrm{Anc}_G(\{i\} \cup S)$, contradicting the orientation $i \ast\!\!\to j$ in H. Therefore, there must be a σ-inducing path given S between i and j in G that is into j. This shows (ii).

(iii) This follows from case (ii) in combination with Lemma 12.2.6.

□

We will frequently use that every mixed graph (of a certain type) that represents a CDMG must be a valid σ-PAG.

Proposition 12.3.7. *Let $H = (J, V, E)$ be a mixed graph with input nodes J, output nodes V and edges E of the types $\{\to, \leftarrow, \leftarrow\!\circ, \leftrightarrow, \circ\!\!-\!\!\circ, \circ\!\!\to, -\!\!\circ, \circ\!\!-, -\}$. Let G be a CDMG with input nodes J and output nodes $V^+ = V \dot{\cup} S$. If H represents G given S, then H is a σ-PAG.*

Proof. By assumption: there is at most one edge between two distinct nodes; there are no edges between a node and itself; no input node is adjacent to any other input node.

We show that H is σ-ancestral. First, suppose that H contained a directed path $i = v_0 \to \ldots \to v_n = j$. By Lemma 12.3.6, then $i \in \mathrm{Anc}_G(\{j\} \cup S)$. If H contained an edge $j \ast\!\!\to i$, this would imply $i \notin \mathrm{Anc}_G(\{j\} \cup S)$, a contradiction. Hence such edges cannot occur. This means that no directed cycles and no almost directed cycles occur in H. It remains to show that no unshielded triple of the form $i \ast\!\!\to j - k$ can occur in H. We prove this by contradiction. Since $j \in \mathrm{Anc}_G(\{k\} \cup S)$ but $j \notin \mathrm{Anc}_G(\{i\} \cup S)$, we must have $j \in \mathrm{Anc}_G(k)$. Also, $k \in \mathrm{Anc}_G(\{j\} \cup S)$. Since $j \in \mathrm{Anc}_G(k)$ and $j \notin \mathrm{Anc}_G(S)$, we must have $k \in \mathrm{Anc}_G(j)$. But then $j \in \mathrm{Sc}_G(k)$, which gives a contradiction with Lemma 12.2.7.

We continue to show that H is maximal. Suppose there is a definitely inducing path μ in H between two distinct nodes $u, v \in V \cup J$. We first show that this implies that $\{u, v\} \not\subseteq J$. If $u, v \in J$, then a definitely inducing path between them cannot consist of a single edge, because all node pairs in J are non-adjacent in H by assumption. Hence it must be of the form $u \ast\!\!\to w_1 \ldots w_n \leftarrow\!\!\ast v$ (with $n \geq 1$) with $w_1, w_n \in V$. But since all nodes $w_1, \ldots, w_n$ are definite colliders, none of them is in $\mathrm{Anc}_G(S)$. Since all nodes $w_1, \ldots, w_n$ are in $\mathrm{Anc}_H(\{u, v\})$, by Lemma 12.3.6, they must all be in $\mathrm{Anc}_G(\{u, v\})$. Since G is a CDMG, they must all be in $\{u, v\} \subseteq J$, a contradiction. Hence, there cannot be a definitely inducing path in H between two nodes in J.

Every edge $i \ast\!\!-\!\!\ast j$ on μ corresponds with a σ-inducing path π_{ij} given S in G between i and j. By Lemma 12.3.6, these σ-inducing paths can be chosen to be into i if the edge is $i \leftarrow\!\!\ast j$, into j if the edge is $i \ast\!\!\to j$, and both into i and j if the edge is $i \leftrightarrow j$. Concatenate all π_{ij} (following the edge ordering of μ) into a walk π in G between u and v. Every non-endpoint node on μ is a definite collider on μ, by assumption. By construction, these nodes then become colliders on π. Since definite colliders on μ are in $\mathrm{Anc}_H(\{u, v\})$ by assumption, they are in $\mathrm{Anc}_G(\{u, v\} \cup S)$ by Lemma 12.3.6. All colliders on some π_{ij} are in $\mathrm{Anc}_G(\{i, j\} \cup S)$. Hence, they are in $\mathrm{Anc}_G(\{u, v\} \cup S)$.

So, all colliders on π are in $\mathrm{Anc}_G(\{u,v\} \cup S)$. All non-endpoint non-colliders on some π_{ij} are unblockable, and therefore all non-endpoint non-colliders on π are unblockable. Therefore, π is a σ-inducing walk given S in G. So there must also be a σ-inducing path given S in G between u and v. Because H represents G given S, we conclude that u, v must be adjacent in H. $\square$

The following result shows that every CDMG can be represented by a σ-PAG given some set of selection nodes.

Proposition 12.3.8. *Let G be a CDMG with input nodes J and output nodes $V^+ = V \dot\cup S$. There exists a σ-PAG, denoted $\mathrm{PAG}^\sigma(G \mid S)$, that represents G given S.*

Proof. We will construct a mixed graph H with input nodes J and output nodes V as follows. Let two nodes $i, j \in J \cup V$ be adjacent in H if and only if $i \neq j$, $\{i,j\} \not\subseteq J$ and there is a σ-inducing path given S between i, j in G. In that case, orient the edge between i and j in H as follows:

$$\begin{cases} i \;\text{—}\; j & \text{if } i \in \mathrm{Anc}_G(\{j\} \cup S) \text{ and } j \in \mathrm{Anc}_G(\{i\} \cup S), \\ i \rightarrow j & \text{if } i \in \mathrm{Anc}_G(\{j\} \cup S) \text{ and } j \notin \mathrm{Anc}_G(\{i\} \cup S), \\ i \leftarrow j & \text{if } i \notin \mathrm{Anc}_G(\{j\} \cup S) \text{ and } j \in \mathrm{Anc}_G(\{i\} \cup S), \\ i \leftrightarrow j & \text{if } i \notin \mathrm{Anc}_G(\{j\} \cup S) \text{ and } j \notin \mathrm{Anc}_G(\{i\} \cup S). \end{cases}$$

It is obvious by construction that H represents G given S. It is a valid σ-PAG by Proposition 12.3.7. $\square$

The σ-PAG constructed in this way contains no circle edge marks by construction, and is therefore maximally informative about ancestral relations (that is, as informative as a σ-PAG can be).

Proposition 12.3.9. *Let G be a CDMG with input nodes J and output nodes $V^+ = V \dot\cup S$ and H a σ-PAG that represents G given S. If $i \ast\!\!\rightarrow j \ast\!\!-\!\!\ast k$ in H (with i, j, k distinct nodes) and $j \in \mathrm{Sc}_G(k)$, then $i \ast\!\!-\!\!\ast k$ in H and $k \notin \mathrm{Anc}_G(\{i\} \cup S)$. If, additionally, the edge in H between i and j is $i \leftrightarrow j$, then also $i \notin \mathrm{Anc}_G(\{k\} \cup S)$.*

Proof. Since $\mathrm{Sc}_G(k) \supseteq \{j,k\}$, k cannot be in J.

By Lemma 12.3.6, $i \ast\!\!\rightarrow j$ in H implies the existence of a σ-inducing walk between i and j given S in G that is into j. This can be extended by concatenation with a directed path from j to k into a σ-inducing walk between i and k given S in G that is into k. Hence, there must be an edge $i \ast\!\!-\!\!\ast k$ in H.

If $k \in \mathrm{Anc}_G(\{i\} \cup S)$, then also $j \in \mathrm{Anc}_G(\{i\} \cup S)$ because $j \in \mathrm{Anc}_G(k)$, a contradiction.

If $i \leftrightarrow j$ in H, then $i \notin \mathrm{Anc}_G(\{j\} \cup S)$. If $i \in \mathrm{Anc}_G(\{k\} \cup S)$, then $i \in \mathrm{Anc}_G(k)$ because $i \notin \mathrm{Anc}_G(S)$; hence also $i \in \mathrm{Anc}_G(k) = \mathrm{Anc}_G(j)$, a contradiction. $\square$

Note that this implies that if $i \ast\!\!\rightarrow j \;\text{—}\; k$ in H, then i and k must be adjacent.

Proposition 12.3.11 states an important result that allows one to deduce the existence of a σ-open path in a CDMG from the existence of certain paths in a σ-PAG that represents the CDMG. In order to state it, we introduce the following notion:

Definition 12.3.10. *Let G be a CDMG with input nodes J and output nodes $V^+ = V \dot\cup S$ and H a σ-PAG that represents G given S. A walk (path) $v_0 \ast\!\!-\!\!\ast \ldots \ast\!\!-\!\!\ast v_n$ in H is called **definitely** m**-open given** $Z \subseteq V \cup J$ if*

1. *every definite collider on the walk is in $\mathrm{Anc}_H(Z)$, and*
2. *every definite non-collider on the walk is not in Z, and*
3. *every node on the walk that is neither a definite collider nor a definite non-collider is in $\mathrm{Anc}_H(Z)/Z$, and*
4. *the walk does not contain any subwalk that could be oriented into the form $v_{i-1} \ast\!\!\rightarrow v_i — v_{i+1}$ or into the form $v_{i-1} — v_i \leftarrow\!\!\ast v_{i+1}$ such that the resulting σ-PAG still represents G given S.*

In certain cases, the existence of a definitely m-open path in a σ-PAG representing a CDMG implies the existence of a corresponding σ-open path in the CDMG.

Proposition 12.3.11. *Let G be a CDMG with input nodes J and output nodes $V^+ = V \dot\cup S$ and H a σ-PAG that represents G given S. Let, for $n \geq 2$,*

$$\pi = q_0 \ast\!\!-\!\!\ast q_1 \ast\!\!\rightarrow q_2 \leftrightarrow \ldots \leftrightarrow q_{n-1} \leftarrow\!\!\ast q_n$$

be a definitely Z-m-open path in H, for $Z \subseteq (J \cup V) \setminus \{q_0, q_n\}$, such that $q_2, \ldots, q_{n-1}$ are definite colliders, and q_1 is either a definite non-collider or a definite collider. Then there exists a $(Z \cup S)$-σ-open path in G between q_0 and q_n.

Proof. The nodes $q_0, q_1, \ldots, q_n$ in $J \cup V$ must all be distinct. For $i = 1, \ldots, n$, let μ_i be a σ-inducing path in G between q_{i-1} and q_i that is into q_{i-1} if the edge $q_{i-1} \leftarrow\!\!\ast q_i$ on π, and into q_i if the edge $q_{i-1} \ast\!\!\rightarrow q_i$ on π (see Lemma 12.3.6). These can be concatenated into a walk $\mu = (\mu_1, \ldots, \mu_n)$ in G between q_0 and q_n:

$$\overbrace{q_0 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast q_1}^{\mu_1}\!\!\!\!\underbrace{\ast\!\!-\!\!\ast \ldots \ast\!\!\rightarrow q_2}_{\mu_2} \leftarrow\!\!\ast \cdots\cdots \ast\!\!\rightarrow \underbrace{q_{n-2} \leftarrow\!\!\ast \ldots \ast\!\!\rightarrow q_{n-1}}_{\mu_{n-1}}\!\!\!\!\overbrace{\leftarrow\!\!\ast \cdots \ast\!\!-\!\!\ast q_n}^{\mu_n}$$

Let $Z \subseteq (J \cup V) \setminus \{q_0, q_n\}$. Since π is assumed to be definitely Z-m-open, $q_2, \ldots, q_{n-1}$ are all in $\mathrm{Anc}_H(Z)$. We must have that $q_1 \in \mathrm{Anc}_G(\{q_0, q_2\} \cup Z \cup S)$, as can be seen by considering the two mutually exclusive cases:

- q_1 is a definite collider on π. Then $q_1 \in \mathrm{Anc}_H(Z)$. By Lemma 12.3.6, $q_1 \in \mathrm{Anc}_G(Z \cup S)$.
- q_1 is a definite non-collider on π. Then $q_1 \notin Z$, but $q_1 \in \mathrm{Anc}_G(\{q_0, q_2\} \cup S)$ because either $q_0 \ast\!\!\leftarrow q_1$ or $q_1 \rightarrow\!\!\ast q_2$ must be on π.

In any case, $q_1 \in \mathrm{Anc}_G(\{q_0, q_2\} \cup Z \cup S)$.

Consider all walks in G between q_0 and q_n with the property that

1. all colliders on it are in $\mathrm{Anc}_G(\{q_0, q_n\} \cup Z \cup S)$, and
2. each non-endpoint non-collider on it is not in $\{q_0, q_n\} \cup Z \cup S$ or is unblockable.

Such walks exist, since the concatenation $\mu = (\mu_1, \dots, \mu_n)$ is one, as we will now show.

To show that the first property holds for μ, note that $q_2, \dots, q_{n-1}$ are in $\mathrm{Anc}_H(Z)$ and hence, by Lemma 12.3.6, $q_2, \dots, q_{n-1} \in \mathrm{Anc}_G(Z \cup S)$, a subset of $\mathrm{Anc}_G(\{q_0, q_n\} \cup Z \cup S)$. We already saw that $q_1 \in \mathrm{Anc}_G(\{q_0, q_n\} \cup Z \cup S)$, which holds in particular if q_1 is a collider on μ. Every internal collider on some μ_i is in $\mathrm{Anc}_G(\{q_{i-1}, q_i\} \cup S)$, and hence also these 'internal' colliders on μ are in $\mathrm{Anc}_G(\{q_0, q_n\} \cup Z \cup S)$.

For the second property, note that all non-endpoint non-colliders on some μ_i are unblockable by assumption. $q_2, \dots, q_{n-1}$ cannot be non-colliders on μ. If q_1 is a non-collider on μ, then it must be a definite non-collider on π and hence $q_1 \notin Z$ (and by assumption, $q_1 \notin S$). Hence, the second property also holds for μ.

Let ν be a walk satisfying both properties above with a minimal number of colliders. We show that all colliders on ν must be in $\mathrm{Anc}_G(Z \cup S)$.

Suppose on the contrary the existence of a collider k on ν that is not in $\mathrm{Anc}_G(Z \cup S)$. It must then be in $\mathrm{Anc}_G(\{q_0, q_n\})$, by assumption. Then there exists a directed path in G from k to q_0 that does not pass through q_n, or there exists a directed path from k to q_n that does not pass through q_0. Without loss of generality, assume the former: there is a directed path in G from k to q_0 in G that does not pass through $Z \cup S \cup \{q_n\}$. Let π be a shortest path of that type. The directed path π from k to q_0 can be concatenated with the subwalk of ν between k and q_n (which is into k) into a walk between q_0 and q_n. This walk has the property, but has fewer colliders than ν: a contradiction.

Therefore, ν is a $(Z \cup S)$-σ-open walk in G between q_0 and q_n. This means that there must also exist a $(Z \cup S)$-σ-open path in G between q_0 and q_n. $\square$

12.4. Unshielded triples

One of the key steps in the FCI algorithm is the orientation of "unshielded triples". The following proposition will later be used to "orient" the edges in unshielded triples in a σ-PAG representing a CDMG.

Proposition 12.4.1. *Let G be a CDMG with input nodes J and output nodes $V^+ = V \dot{\cup} S$ and H a σ-PAG that represents G given S. If (i, j, k) form an unshielded triple in H with $i \in V$ (and $j, k \in V \cup J$), then either*

(i) $j \notin Z$ for each $Z \subseteq (V \cup J) \setminus \{i, k\}$ such that $i \perp_G^\sigma k \mid Z \cup S$, and $j \notin \mathrm{Anc}_G(\{i, k\} \cup S)$, or

(ii) $j \in Z$ for each $Z \subseteq (V \cup J) \setminus \{i, k\}$ such that $i \perp_G^\sigma k \mid Z \cup S$, and $j \in \mathrm{Anc}_G(\{i, k\} \cup S)$.

Proof. Case (i): $j \notin \mathrm{Anc}_G(\{i, k\} \cup S)$. By orienting the edge marks at j on the path $i \ast\!\!-\!\!\ast\, j \ast\!\!-\!\!\ast\, k$ in H into $i \ast\!\!\rightarrow j \leftarrow\!\!\ast\, k$ (if not already oriented that way), we obtain a σ-PAG $\tilde{H}$ that represents G. Let $Z \subseteq (V \cup J) \setminus \{i, k\}$ such that $i \perp_G^\sigma k \mid Z \cup S$. If $j \in Z$, the path $i \ast\!\!\rightarrow j \leftarrow\!\!\ast\, k$ would be definitely Z-m-open in $\tilde{H}$. By Proposition 12.3.11, this

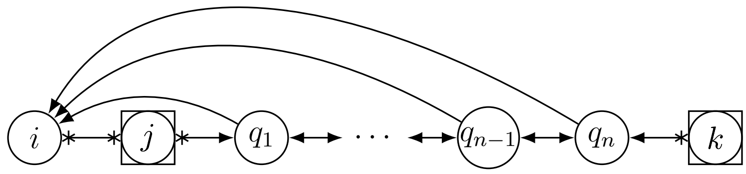


Figure 33: Discriminating path $(i, j, q_1, \dots, q_n, k)$ for j between i and k. Only j and k could be an input node, all other nodes must be output nodes.

implies the existence of a $(Z \cup S)$-σ-open walk in G between i and k, a contradiction. Hence $j \notin Z$.

Case (ii): $j \in \mathrm{Anc}_G(\{i,k\} \cup S)$. By orienting the edge marks at j on the path $i \ast\!\!\rightarrow j \leftarrow\!\!\ast k$ in H into $i \ast\!\!- j \leftarrow\!\!\ast k$ (if $j \in \mathrm{Anc}_G(\{i\} \cup S)$) or $i \ast\!\!\rightarrow j -\!\!\ast k$ (if $j \in \mathrm{Anc}_G(\{k\} \cup S)$) or $i \ast\!\!- j -\!\!\ast k$ (if $j \in \mathrm{Anc}_G(\{i\} \cup S) \cap \mathrm{Anc}_G(\{k\} \cup S)$), we obtain a σ-PAG $\tilde{H}$ that represents G. Let $Z \subseteq (V \cup J) \setminus \{i,k\}$ such that $i \perp^{\sigma}_{G} k \mid Z \cup S$. Note that with Proposition 12.3.9, we can exclude the possibility that the path $i \ast\!\!\rightarrow j \leftarrow\!\!\ast k$ can be oriented into $i — j \leftarrow\!\!\ast k$ or $i \ast\!\!\rightarrow j — k$ such that the resulting σ-PAG still represents G given S, because i and k are not adjacent by assumption. Hence, if $j \notin Z$, the path $i \ast\!\!\rightarrow j \leftarrow\!\!\ast k$ must be definitely Z-m-open in $\tilde{H}$. By Proposition 12.3.11, this implies the existence of a $(Z \cup S)$-σ-open walk in G between i and k, a contradiction. Hence $j \in Z$. $\square$

Note that in the first case, we can orient the edges as $i \ast\!\!\rightarrow j \leftarrow\!\!\ast k$ (if they were not already oriented in this way) to obtain a σ-PAG $\tilde{H}$ that also represents G. In the second case, we cannot orient the edges, since we don't know whether $j \in \mathrm{Anc}_G(\{i\} \cup S)$ or $j \in \mathrm{Anc}_G(\{k\} \cup S)$ or both.

12.5. Discriminating paths

Another step in the FCI algorithm is related to the notion of "discriminating paths". This can be considered as an extension of the notion of unshielded triple.

Definition 12.5.1. *A path $\pi = (i, j, q_1, \dots, q_n, k)$ (with $n \geq 1$) in a mixed graph H is a* ***discriminating path for*** j *if:*

(i) i is not adjacent to k in H, and

(ii) for $r = 1, \dots, n$: q_r is a definite collider on π and a parent of i in H.

Figure 33 illustrates this notion.

Remark 12.5.2. *It is instructive to think about a discriminating path rather as a certain collection of paths:*

$$
\begin{aligned}
&i \leftarrow q_n \leftarrow\!\!\ast\; k \\
&i \leftarrow q_{n-1} \leftrightarrow q_n \leftarrow\!\!\ast\; k \\
&i \leftarrow q_{n-2} \leftrightarrow q_{n-1} \leftrightarrow q_n \leftarrow\!\!\ast\; k \\
&\vdots \\
&i \leftarrow q_1 \leftrightarrow \ldots \leftrightarrow q_{n-1} \leftrightarrow q_n \leftarrow\!\!\ast\; k \\
&i \;\ast\!\!-\!\!\ast\; j \;\ast\!\!\rightarrow q_1 \leftrightarrow \ldots \leftrightarrow q_{n-1} \leftrightarrow q_n \leftarrow\!\!\ast\; k
\end{aligned}
$$

with the additional requirement that i and k are not adjacent.

The following quintessential property of discriminating paths is analogous to that of unshielded triples.

Proposition 12.5.3. *Let G be a CDMG with input nodes J and output nodes $V^+ = V \dot{\cup} S$ and H a σ-PAG that represents G given S. If $(i, j, q_1, \ldots, q_n, k)$ is a discriminating path in H for j between i and k, then either*

(i) $j \notin Z$ for each $Z \subseteq (V \cup J) \setminus \{i, k\}$ such that $i \perp_G^\sigma k | Z \cup S$, and $j \notin \mathrm{Anc}_G(\{q_1, i\} \cup S)$, or

(ii) $j \in Z$ for each $Z \subseteq (V \cup J) \setminus \{i, k\}$ such that $i \perp_G^\sigma k | Z \cup S$, and $j \in \mathrm{Anc}_G(\{i\} \cup S)$.

In both cases, $i \notin \mathrm{Anc}_G(\{j\} \cup S)$.

Proof. Since q_1 is a parent of i in H, $q_1 \in \mathrm{Anc}_G(\{i\} \cup S)$. Because q_1 is a definite collider in H, $q_1 \notin \mathrm{Anc}_G(S)$. Hence $q_1 \in \mathrm{Anc}_G(i) \setminus \{i\}$, which means that $i \in V$. Also, this implies that $i \notin \mathrm{Anc}_G(j)$; otherwise, $q_1 \in \mathrm{Anc}_G(j)$ which contradicts $j \;\ast\!\!\rightarrow q_1$ in H. Hence $i \notin \mathrm{Anc}_G(\{j\} \cup S)$. By orienting the edge $i \;\ast\!\!-\!\!\ast\; j$ in H as $i \leftarrow\!\!\ast\; j$, we obtain a σ-PAG $\tilde{H}$ that represents G.

We first show that if $Z \subseteq (V \cup J) \setminus \{i, k\}$ such that $i \perp_G^\sigma k \,|\, Z \cup S$, then $\{q_1, \ldots, q_n\} \in Z$. This can be seen from the various subpaths in Remark 12.5.2. From the first path: if $q_n \notin Z$ then this path would be definitely m-open in $\tilde{H}$, and hence (by Proposition 12.3.11) there must be a $(Z \cup S)$-σ-open walk in G between i and k, which would contradict $i \perp_G^\sigma k \,|\, Z \cup S$. Once we have shown that $\{q_n, q_{n-1}, \ldots, q_{n-r+1}\} \in Z$ for $1 \le r < n$, we see from the $r + 1$'th path that $q_{n-r} \in Z$ to avoid definitely m-opening up this path in $\tilde{H}$, and thereby avoiding the existence of a $(Z \cup S)$-σ-open path in G between i and k.

Case (i): $j \notin \mathrm{Anc}_G(\{q_1, i\} \cup S)$. By orienting the edge marks at j on the path $i \leftarrow\!\!\ast\; j \;\ast\!\!-\!\!\ast\; q_1$ in $\tilde{H}$ into $i \leftrightarrow j \leftarrow\!\!\ast\; q_1$ (if not already oriented that way), we obtain a σ-PAG $\bar{H}$ that represents G. Let $Z \subseteq (V \cup J) \setminus \{i, k\}$ such that $i \perp_G^\sigma k \,|\, Z \cup S$. If $j \in Z$, the discriminating path in $\bar{H}$ is definitely Z-m-open. By Proposition 12.3.11, this implies the existence of a $(Z \cup S)$-σ-open walk in G between i and k. Contradiction. Hence $j \notin Z$.

Case (ii): $j \in \mathrm{Anc}_G(\{q_1, i\} \cup S)$. Since $q_1 \in \mathrm{Anc}_G(i)$, this implies $j \in \mathrm{Anc}_G(\{i\} \cup S)$. By orienting the the edge $i \leftarrow\!\!* \, j$ in $\tilde{H}$ as $i \leftarrow j$, we obtain a σ-PAG $\bar{H}$ that represents G. Let $Z \subseteq (V \cup J) \setminus \{i, k\}$ such that $i \perp^{\sigma}_{G} k \mid Z \cup S$. If $j \notin Z$, the discriminating path in $\bar{H}$ is definitely Z-m-open. By Proposition 12.3.11, this implies the existence of a $(Z \cup S)$-σ-open walk in G between i and k. Contradiction. Hence $j \in Z$. □

Note that in the first case, we can orient $i \leftrightarrow j \leftrightarrow q_1$ (as far as the edge marks were not already oriented in this way) to obtain a σ-PAG $\bar{H}$ that also represents G. In the second case, we can orient $i \leftarrow j$ (as far as the edge marks were not already oriented in this way) to obtain a σ-PAG $\bar{H}$ that also represents G.

12.6. Independence models and Markov equivalence

We start with an abstract definition of an "independence model", where we extend the common definition to allow for input nodes.

Definition 12.6.1. *Given two disjoint sets J, V (the 'inputs' and 'outputs', respectively), we call a subset of*

$$\{(A, B, C) : A, B, C \subseteq J \cup V, A \cap J = \emptyset, J \subseteq B \cup C\}$$

an ***independence model over*** *$V \mid J$. For an element (A, B, C) of an independence model, we also say that C* ***separates*** *A* ***from*** *B.*

Independence models can be used to encode all (conditional) independences in a Markov kernel.

Definition 12.6.2. *For a Markov kernel $K(X_V|X_J) : \mathcal{X}_J \dashrightarrow \mathcal{X}_V$, we define its* ***independence model*** *to be*

$$\mathrm{IM}(K(X_V|X_J)) := \{(A, B, C) : A, B, C \subseteq J \cup V, A \cap J = \emptyset, J \subseteq B \cup C : X_A \underset{K(X_V|X_J)}{\perp\!\!\!\perp} X_B \mid X_C\},$$

i.e., the set of all conditional independences (of restricted form) that $K(X_V|X_J)$ satisfies.

Independence models can also encode all separations in a graph.

Definition 12.6.3. *For a CDMG G with input nodes J and output nodes V, define its* ***σ-independence model*** *to be*

$$\mathrm{IM}_\sigma(G) := \{(A, B, C) : A, B, C \subseteq J \cup V, A \cap J = \emptyset, J \subseteq B \cup C : A \overset{\sigma}{\underset{G}{\perp}} B \mid C\},$$

i.e., the set of all σ-separations (of restricted form) entailed by the graph. Define its ***d-independence model*** *to be*

$$\mathrm{IM}_d(G) := \{(A, B, C) : A, B, C \subseteq J \cup V, A \cap J = \emptyset, J \subseteq B \cup C : A \overset{d}{\underset{G}{\perp}} B \mid C\},$$

i.e., the set of all d-separations (of restricted form) entailed by the graph.

Both $\mathrm{IM}_d(G)$ and $\mathrm{IM}_\sigma(G)$ are independence models over $V \mid J$ (with J the input nodes of G and V the output nodes of G). For CADMGs, σ-separation is equivalent to d-separation, and hence, if G is acyclic, then $\mathrm{IM}_d(G) = \mathrm{IM}_\sigma(G)$.

When conditioning on a set of selection variables, we can also define conditional independence models from graphs as follows.

Definition 12.6.4. *For a CDMG G with input nodes J and output nodes $V^+ = V \dot\cup S$, define its* ***σ-independence model given*** *S to be*

$$\mathrm{IM}_\sigma(G \mid S) := \{(A,B,C) : A,B,C \subseteq J \cup V, A \cap J = \emptyset, J \subseteq B \cup C : A \underset{G}{\overset{\sigma}{\perp}} B \mid C \cup S\},$$

i.e., the set of all σ-separations (of restricted form) involving nodes not in S, when also conditioning on S. Define its ***d-independence model given*** *S to be*

$$\mathrm{IM}_d(G \mid S) := \{(A,B,C) : A,B,C \subseteq J \cup V, A \cap J = \emptyset, J \subseteq B \cup C : A \underset{G}{\overset{d}{\perp}} B \mid C \cup S\},$$

i.e., the set of all d-separations (of restricted form) involving nodes not in S, when also conditioning on S.

Hence, also $\mathrm{IM}_\sigma(G \mid S)$ and $\mathrm{IM}_d(G \mid S)$ are independence models over $V \mid J$ (with J the input nodes of G and V the output nodes of G except for those in S). For CADMGs, σ-separation is equivalent to d-separation, and hence, if G is acyclic, then $\mathrm{IM}_d(G \mid S) = \mathrm{IM}_\sigma(G \mid S)$.

For notational convenience, we define symmetrized versions of independence models.

Definition 12.6.5. *Given two disjoint sets J, V (the 'inputs' and 'outputs', respectively), and an independence model I over $V \mid J$, we define the* ***symmetrized*** *independence model as*

$$I^+ := I \cup \{(B,A,C) : (A,B,C) \in I\}.$$

The input to the extended FCI algorithm will consist of an independence model over $V \mid J$, and its output will consist of a mixed graph with input nodes J and output nodes V. One of the nice properties of the extended FCI algorithm is that if the input of FCI is the independence model of a CDMG given S, then the output of FCI will be a σ-PAG that represents the "Markov equivalence class" of the CDMG (at least if $S = \emptyset$ or $J = \emptyset$).

Definition 12.6.6. *Let G_1, G_2 be two CDMGs with input nodes J and output nodes $V_1^+ = V \dot\cup S_1$, $V_2^+ = V \dot\cup S_2$, respectively. We call G_1 and G_2* ***σ-Markov equivalent w.r.t.*** *$V \mid J$ if $\mathrm{IM}_\sigma(G_1 \mid S_1) = \mathrm{IM}_\sigma(G_2 \mid S_2)$, and* ***$d$-Markov equivalent w.r.t.*** *$V \mid J$ if $\mathrm{IM}_d(G_1 \mid S_1) = \mathrm{IM}_d(G_2 \mid S_2)$.*

When exploiting conditional independences for causal discovery, one typically has to make some kind of faithfulness assumption. The faithfulness assumption that we will make for the extended FCI algorithm will be that the (restricted) conditional independences in the conditioned Markov kernel are identical to the (restricted) σ-separations in the graph, given S:

$$\mathrm{IM}(P_M(X_V \mid X_S \in \xi_S, \mathrm{do}(X_J))) = \mathrm{IM}_\sigma(G_{V^+|J}(M) \mid S).$$

12.7. Skeleton search

The FCI algorithm consists of two phases, the skeleton search phase, which is followed by the orientation phase. In this subsection, we describe the skeleton search phase.

Definition 12.7.1. *Given a σ-PAG $H = (J, V, E)$, its* ***skeleton*** *is the mixed graph* $\mathrm{skel}(H) := (J, V, F)$ *with the same nodes, and with a bicircle edge $i \circ\!\!-\!\!\circ j$ in F if and only if $i *\!\!-\!\!* j$ in E (i.e., if i and j are adjacent in H).*

Hence, the only edge type occurring in the skeleton is the bicircle edge. The skeleton has no edge between any pair of input nodes. We will later frequently refer to the unordered pairs of nodes that may be adjacent:

Definition 12.7.2. *For input nodes J and output nodes V, define*

$$\mathtt{separable}(V|J) := \{\{i, j\} : i \in V, j \in J \cup V, i \neq j\}.$$

The aim of the skeleton search phase of the FCI algorithm is to construct the skeleton of the σ-PAG that represents a CDMG given S from the σ-independence model given S of the CDMG. It does this by testing for each separable edge in the skeleton whether it can find any subset of nodes that separates the two nodes. If it finds a separating set between an (unordered) pair of distinct nodes $\{i, j\}$, the set is stored as $\mathtt{sepset}(\{i, j\})$. The orientation phase later makes use of these separating sets found in the skeleton phase.

A brute-force search over all possible subsets of $(J \cup V) \setminus \{i, j\}$, as in Algorithm 1, would be a straightforward solution. However, it is also computationally extremely expensive for all but the smallest cardinalities of V and J, and statistically not very reliable in case the separations have to be tested with conditional independence tests on finite data.

Algorithm 1 Brute-force skeleton algorithm.

input Input node set J; output node set V; independence model I over $V \mid J$
output mixed graph H with input nodes J and output nodes V; separating sets $\mathtt{sepset}$
$H \leftarrow (J, V, \emptyset)$
for each $(i, j) \in \mathtt{separable}(V|J)$ **do**
    add edge $i \circ\!\!-\!\!\circ j$ to H
    for each $Z \subseteq (V \cup J) \setminus \{i, j\}$ **do**
        if $(i, j, Z) \in I^+$ **then** ▷ found a separating set
            delete edge $i \circ\!\!-\!\!\circ j$ from H
            $\mathtt{sepset}(\{i, j\}) \leftarrow Z$
            break
        end if
    end for
end for

To get some inspiration on how to address this, we will first describe the skeleton search phase of the PC algorithm, an ancestor of the FCI algorithm designed for DAGs

[SGS00]. The PC skeleton phase (Algorithm 2) searches for separating sets of increasing cardinality. As candidates for a separating set between nodes $i, j \in H$, it considers all subsets of $\mathrm{Adj}_H(i)$ and all subsets of $\mathrm{Adj}_H(j)$. The rationale is that if G is a DAG, then either $\mathrm{Pa}_G(i) \subseteq \mathrm{Adj}_H(i)$ or $\mathrm{Pa}_G(j) \subseteq \mathrm{Adj}_H(j)$ d-separates i from j.

Algorithm 2 Original PC skeleton algorithm.

input Output node set V; independence model I over $V \mid \emptyset$
output mixed graph H with output nodes V; separating sets `sepset`
initialize H as a complete graph with only bicircle ($\circ\!\!-\!\!\circ$) edges
$n \leftarrow 0$
repeat
    repeat
        select $i, j \in V$ with $i \circ\!\!-\!\!\circ j$ in H and $\#(\mathrm{Adj}_H(i) \setminus \{j\}) \geq n$
        select a subset $Z \subseteq \mathrm{Adj}_H(i) \setminus \{j\}$ of cardinality n
        if $(i, j, Z) \in I^+$ **then**
            delete edge $i \circ\!\!-\!\!\circ j$ from H
            `sepset`$(\{i, j\}) \leftarrow Z$
        end if
    until no more such tuples (i, j, Z) can be selected
    $n \leftarrow n + 1$
until for all $i, j \in V$ with $i \circ\!\!-\!\!\circ j$ in H, $\#(\mathrm{Adj}_H(i) \setminus \{j\}) < n$

We can easily extend this to deal with input nodes as well, see Algorithm 3. We may restrict ourselves to search for separating sets that contain all nodes in J (except j itself, if $j \in J$). We therefore only need to consider subsets of neighbours of i that are not in J, in increasing cardinality.

The underlying idea may no longer hold if G is a CADMG or a CDMG, or in case of selection bias. The original proposal (which motivated the somewhat optimistic adjective "Fast" in the name of the FCI algorithm [SMR95]) replaces the subsets of adjacent nodes by so-called "Possible-D-Sep" sets.[85] Before we can define these, we need some definitions and theory.

Definition 12.7.3. *Let G be a CDMG with input nodes J and output nodes $V^+ = V \dot\cup S$. For distinct nodes $i, j \in V \dot\cup J$, define $\mathtt{SEP}_G(i, j)$ as the set of nodes $k \in V$ such that $k \neq i$ and there is a walk π in G between i and k such that every node on π is in $\mathrm{Anc}_G(\{i, j\} \cup S)$, and every non-endpoint non-collider on π is unblockable.*

The name $\mathtt{SEP}_G(i, j)$ is motivated by the following property.

Proposition 12.7.4. *Let G be a CDMG with input nodes J and output nodes $V^+ = V \dot\cup S$. Let $i \in V$ and $j \in J \cup V$ be distinct nodes. If there exists no σ-inducing walk given S between i, j in G, then $\mathtt{SEP}_G(i, j) \cap \{i, j\} = \emptyset$ and $i \perp_G^\sigma j \mid \mathtt{SEP}_G(i, j) \cup (J \setminus \{j\}) \cup S$.*

[85] An alternative search strategy was proposed that can be considerably faster in practice, for which it can be shown that the corresponding FCI+ algorithm is of polynomial-time complexity in the number of nodes, as long as the degree of the DPAG is bounded [CMH13].

Algorithm 3 Extended PC skeleton algorithm $\texttt{PCskeleton}(J, V, I)$.

input Input node set J; output node set V; independence model I over $V \mid J$

output mixed graph H with input nodes J and output nodes V; separating sets $\texttt{sepset}$

$H \leftarrow (J, V, \emptyset)$

for each $(i, j) \in \texttt{separable}(V|J)$ **do**

  add edge $i \circ\!\!-\!\!\circ j$ to H

end for

$n \leftarrow 0$

repeat

  repeat

    select $i \in V, j \in V \cup J$ with $i \circ\!\!-\!\!\circ j$ in H and $\#(\mathrm{Adj}_H(i) \setminus (J \cup \{j\})) \geq n$

    select a subset $Z \subseteq \mathrm{Adj}_H(i) \setminus (J \cup \{j\})$ of cardinality n

    if $(i, j, Z \cup (J \setminus \{j\})) \in I^+$ **then**

      delete edge $i \circ\!\!-\!\!\circ j$ from H

      $\texttt{sepset}(\{i, j\}) \leftarrow Z \cup (J \setminus \{j\})$

    end if

  until no more such tuples (i, j, Z) can be selected

  $n \leftarrow n + 1$

until for all $i \in V, j \in V \cup J$ with $i \circ\!\!-\!\!\circ j$ in H, $\#(\mathrm{Adj}_H(i) \setminus (J \cup \{j\})) < n$

Proof. By definition, $\texttt{SEP}_G(i, j) \subseteq \mathrm{Anc}_G(\{i, j\} \cup S)$ and $i \notin \texttt{SEP}_G(i, j)$. If $j \in \texttt{SEP}_G(i, j)$, the walk π in Definition 12.7.3 between i and j must be σ-inducing given S. Since there exists no σ-inducing walk given S between i, j in G by assumption, $j \notin \texttt{SEP}_G(i, j)$.

We prove the σ-separation by contradiction. Write $Z := \texttt{SEP}_G(i, j) \cup (J \setminus \{j\}) \cup S$. Suppose there exists a path in G between i and $\{j\} \cup J$ that is σ-open given Z. It cannot end in a node in $J \setminus \{j\}$, and hence it must end in j. Let $\pi = v_0 \ast\!\!-\!\!\ast \dots \ast\!\!-\!\!\ast v_n$ with $v_0 = i$ and $v_n \in j$ be such a path consisting of a minimal number of nodes. Every node on π must be in $\mathrm{Anc}_G(\{i, j\} \cup S) \cup J$, since by Lemma 10.5.2, each node on π is in $\mathrm{Anc}_G((\{i, j\} \setminus J) \cup Z)$, and $Z \subseteq \mathrm{Anc}_G(\{i, j\} \cup S) \cup J$. π can only contain nodes in J as endnodes (otherwise we could shorten the path). In other words, the only node on π that could be in J is j.

Denote the subwalk of π from v_a to (and including) v_b by $\pi(v_a, v_b)$, for $0 \leq a \leq b \leq n$. We will show that for all $k = 1, \dots, n$, $\pi(v_0, v_k)$ has the property that every non-endpoint non-collider on it is unblockable. The property trivially holds for $k = 1$. Suppose it holds for $k < n$. Since $v_0, \dots, v_k$ are all in $\mathrm{Anc}_G(\{i, j\} \cup S)$, and all non-endpoint non-colliders on $\pi(v_0, v_k)$ are unblockable, we conclude that $v_k \in \texttt{SEP}_G(i, j)$. If v_k is a non-collider on $\pi(v_0, v_{k+1})$, it must be unblockable, because $\pi(v_0, v_{k+1})$ is Z-σ-open and $v_k \in \texttt{SEP}_G(i, j) \subseteq Z$. So the property also holds for $k + 1$.

In particular, we can conclude that $j = v_n \in \texttt{SEP}_G(i, j)$, a contradiction. □

In practice, one does not know the set $\texttt{SEP}_G(i, j)$ if G is unknown. One can, however, easily obtain a 'bound' on this set by identifying a superset.

Definition 12.7.5. *Let $H_0 = (J, V, E)$ be a mixed graph with input nodes J, output nodes V and edges E of the types $\{\to, \leftarrow, \leftarrow\!\circ, \leftrightarrow, \circ\!-\!\circ, \circ\!\!\to, -\!\circ, \circ\!-, -\}$. For $i \in V, j \in J \cup V$ distinct, we define $\mathtt{posSEP}_{H_0}(i,j) \subseteq V$ to consist of those nodes $k \in V$ such that $k \notin \{i,j\}$ and there is a path between i and k in H_0 such that for every subsequent triple $a \ast\!\!\to b \ast\!\!-\!\!\ast c$ on the path, either the triple is a definite collider in H_0, or a triangle in H_0.*

We can now show that:

Lemma 12.7.6. *Let G be a CDMG with input nodes J and output nodes $V^+ = V \dot{\cup} S$. Let H_0 be the mixed graph constructed by lines 3–8 of Algorithm 4 when run on the independence model $\mathrm{IM}_\sigma(G \mid S)$. Then, for $i \in V$ and $j \in J \cup V$ distinct nodes, if there is no σ-inducing path given S in G between i and j, then $\mathtt{SEP}_G(i,j) \subseteq \mathtt{posSEP}_{H_0}(i,j)$.*

Proof. Note that the skeleton of H_0 is a supergraph of the skeleton of $\mathrm{PAG}^\sigma(G \mid S)$, i.e., every adjacency in $\mathrm{PAG}^\sigma(G \mid S)$ must also be an adjacency in H_0. Let $k \in \mathtt{SEP}_G(i,j)$. Since there is no σ-inducing path given S in G between i and j, $\{i,j\} \cap \mathtt{SEP}_G(i,j) = \emptyset$ by Proposition 12.7.4, and hence $k \neq i$ and $k \neq j$. By definition, there is a path π in G between i and k with the property that every node on π is in $\mathrm{Anc}_G(\{i,j\} \cup S) \cap V$, and every non-endpoint non-collider on π is unblockable. There is a corresponding path π' in H_0 between i and k that consists of the same sequence of nodes (but may have different edges between the nodes).

Consider a subsequent triple $a \ast\!\!\to b \ast\!\!-\!\!\ast c$ on π. Suppose b is a collider on π. In H_0, a must also be adjacent to b, and b to c. If a and c are not adjacent in H_0, then a separating set for a and c was found during the construction of H_0, or a and c are both in J. The latter would be a contradiction. Therefore, it will have been oriented as a definite collider in H_0 according to Proposition 12.4.1. Otherwise (if a and c are adjacent in H_0), it forms a triangle in H_0. If b is a non-collider on π, then it must be unblockable. But in that case, $a \ast\!\!\to b \ast\!\!\to c$ is a σ-inducing walk given S between a and c in G. Hence, (a,b,c) will form a triangle in H_0. Therefore, $k \in \mathtt{posSEP}_{H_0}(i,j)$. $\square$

So we do not need to search over all possible subsets of $(J \cup V) \setminus \{i,j\}$ for a separating set between i, j, but only the subsets in $\mathtt{posSEP}_{H_0}(i,j)$. The skeleton phase of the extended FCI algorithm is described in Algorithm 4. It is an extension of the original FCI skeleton search [SMR95] to deal with input nodes. It first runs the extended PC skeleton phase (Algorithm 3) and orients unshielded triples, obtaining a directed mixed graph H_0. Then it calculates the sets $\mathtt{posSEP}_{H_0}(i,j)$ for all distinct pairs (i,j) with $i \in V, j \in J \cup V$. If $i \perp^\sigma_G j \mid Z \cup S$ for some $Z \subseteq (J \cup V) \setminus \{i,j\}$, then $i \perp^\sigma_G j \mid Z^* \cup S$ for some $Z^* \subseteq \mathtt{posSEP}_{H_0}(i,j)$. Since some of the oriented colliders may be incorrect in H_0 (because the PC skeleton phase may not have found all separating sets), it removes all orientations from H_0 and then continues with a more extensive search for separating sets, similar to how the PC skeleton phase is done, but now using $\mathtt{posSEP}_{H_0}(i,j)$ instead of $\mathrm{Adj}_{H_0}(i) \setminus \{j\}$ to find candidate nodes for the separating set.

Theorem 12.7.7. *The extended FCI skeleton algorithm (Algorithm 4) is **sound**: if its input consists of the σ-independence model $\mathrm{IM}_\sigma(G \mid S)$ given S of a CDMG G, then its*

Algorithm 4 Extended FCI skeleton algorithm `FCIskeleton`(J, V, I).

input Input node set J; output node set V; independence model I over $V \mid J$
output mixed graph H with input nodes J and output nodes V; separating sets `sepset`
$(H_0, \texttt{sepset}) \leftarrow \texttt{PCskeleton}(J, V, I)$
for each unshielded triple (i, j, k) in H_0 with $i, j \in V$ **do** ▷ orient colliders
    if $j \notin \texttt{sepset}(\{i, k\})$ **then**
        orient it as $i \ast\!\!\rightarrow j \leftarrow\!\!\ast k$
    end if
end for
for each $i \in V, j \in V \cup J$ with $i \neq j$ **do** ▷ construct Possible-D-Sep sets
    calculate $\texttt{posSEP}_{H_0}(i, j)$
end for
$H \leftarrow (J, V, \{i \circ\!\!-\!\!\circ j : i \ast\!\!-\!\!\ast j \in H_0\})$ ▷ forget orientations
$n \leftarrow 0$
repeat ▷ search for separating sets
    repeat
        select $i \in V, j \in V \cup J$ with $i \circ\!\!-\!\!\circ j$ in H and $\#(\texttt{posSEP}_{H_0}(i, j)) \geq n$
        select a subset $Z \subseteq \texttt{posSEP}_{H_0}(i, j)$ of cardinality n
        if $(i, j, Z \cup (J \setminus \{j\})) \in I$ **then**
            delete edge $i \circ\!\!-\!\!\circ j$ from H
            $\texttt{sepset}(\{i, j\}) \leftarrow Z \cup (J \setminus \{j\})$
        end if
    until no more such tuples (i, j, Z) can be selected
    $n \leftarrow n + 1$
until for all $i \in V, j \in V \cup J$ with $i \circ\!\!-\!\!\circ j$ in H, $\#(\texttt{posSEP}_{H_0}(i, j)) < n$

output will be $\mathrm{skel}(\mathrm{PAG}^\sigma(G\,|\,S))$. *Furthermore,* $i \perp^\sigma j \,|\, \mathtt{sepset}(\{i,j\})$ *for all* $i \in V, j \in V \cup J$ *for node pairs* $(i,j) \in \mathtt{separable}(V|J)$ *that are non-adjacent in the skeleton.*

Proof. Let G be a CDMG with input nodes J and output nodes $V^+ = V \dot\cup S$ and $I = \mathrm{IM}_\sigma(G\,|\,S)$ its σ-independence model given S. By Lemma 12.7.6, the mixed graph H_0 constructed by lines 3–8 has the property that for distinct $i \in V, j \in J \cup V$, if there is no σ-inducing path given S in G between i and j, then $\mathtt{SEP}_G(i,j) \subseteq \mathtt{posSEP}_{H_0}(i,j)$. Thus, if $i \in V$ and $j \in J \cup V$ are not adjacent in $\mathrm{skel}(\mathrm{PAG}^\sigma(G\,|\,S))$, then we are guaranteed that for some subset $Z \subseteq \mathtt{posSEP}_{H_0}(i,j)$, we have that $i \perp^\sigma_G j \,|(Z \cup J) \setminus \{j\} \cup S$. The graph H constructed in lines 12–24 will be a mixed graph with input nodes J and output nodes V that only has bicircle edges. It has an edge between any pair of distinct nodes $i,j \in J \cup V$ if and only if $\{i,j\} \not\subseteq J$ and there is no set $Z \subseteq (J \cup V) \setminus \{i,j\}$ such that $i \perp^\sigma_G j \,|\, Z \cup S$. Furthermore, if there is no edge in H between a pair of distinct nodes $i,j \in J \cup V$, then either $\{i,j\} \subseteq J$, or $i \perp^\sigma_G j \,|\, \mathtt{sepset}(\{i,j\}) \cup S$. By Proposition 12.2.3, this implies that two distinct nodes $i,j \in J \cup V$ are adjacent in H at this stage if and only if there is a σ-inducing walk given S in G between i and j. Hence H must be $\mathrm{skel}(\mathrm{PAG}^\sigma(G\,|\,S))$. □

12.8. FCI Algorithm

We are now ready to describe a causal inference algorithm that is an extension of the original Fast Causal Inference (FCI) algorithm of [SMR95] to deal with input nodes and cycles. It is presented as Algorithm 5. Its input is an independence model over $V\,|\,J$, where V and J are index sets of output and input nodes, respectively. Its output is a mixed graph with input nodes J and output nodes V. It starts with a *skeleton phase* (line 3, see Algorithm 4) that is aimed at deducing the adjacencies between the nodes, and to find sets that separate two separable node pairs. Then, it runs various *orientation rules* (lines 4–16) that iteratively orient circle edge marks into tails and arrowheads. Note that by convention, the labeled nodes within each orientation rule are assumed to be distinct (for example, in $\mathcal{R}0$, it is implicitly assumed that $i \neq j \neq k \neq i$). For the special case $J = \emptyset$, the algorithm reduces to the standard formulation of the FCI algorithm [Zha08].[86]

Theorem 12.8.1 (Extended FCI Soundness)**.** *The Extended FCI algorithm (Algorithm 5) is* ***sound****: if its input consists of the* σ*-independence model* $\mathrm{IM}_\sigma(G\,|\,S)$ *of a CDMG* G *given* S*, then its output will be a valid* σ*-PAG* H *that represents* G *given* S*.*

Proof. Let G be a CDMG with input nodes J and output nodes $V^+ = V \cup S$ and $I = \mathrm{IM}_\sigma(G\,|\,S)$ its σ-independence model given S.

[86] Compared to the standard formulation of [Zha08], which assumes no input nodes, we have adapted the skeleton search phase (by starting with a mixed graph that contains no edges between input nodes, limiting the paths in the calculation of the $\mathtt{posSEP}_{H_0}(i,j)$ sets to output nodes only, and by including all input nodes, except j itself, into the separating set). Furthermore, we added step 4 to orient the edges between input and output nodes. The formulation of orientation rules $\mathcal{R}0$, $\mathcal{R}1$, $\mathcal{R}3$, $\mathcal{R}7$ is slightly different for this extended version (as these would not be valid in case both $i,k \in J$), and the rest of the algorithm is unchanged.

Algorithm 5 Extended FCI Algorithm.

input Input node set J; output node set V; independence model I over $V \mid J$
output mixed graph H with input nodes J and output nodes V
$(H, \texttt{sepset}) \leftarrow \texttt{FCIskeleton}(J, V, I)$
for each edge $j \circ\!\!-\!\!\circ v$ in H with $j \in J$, $v \in V$ **do**
    orient $j -\!\!\circ v$
end for
repeat

$\mathcal{R}0$ if $i \ast\!\!-\!\!\ast j \ast\!\!-\!\!\ast k$ in H with $i \notin J$, and i and k are not adjacent in H, then orient $i \ast\!\!\rightarrow j \leftarrow\!\!\ast k$ if $j \notin \texttt{sepset}(\{i, k\})$

until this orientation rule is not applicable
repeat

$\mathcal{R}1$ if $i \ast\!\!-\!\!\circ j \leftarrow\!\!\ast k$ in H with $i \notin J$, and i and k are not adjacent in H, then orient $i \leftarrow j$

$\mathcal{R}2$ if $i \rightarrow j \ast\!\!\rightarrow k$ or $i \ast\!\!\rightarrow j \rightarrow k$ in H, and $i \ast\!\!-\!\!\circ k$ in H, then orient $i \ast\!\!\rightarrow k$

$\mathcal{R}3$ if $i \ast\!\!\rightarrow j \leftarrow\!\!\ast k$ and $i \ast\!\!-\!\!\circ l \circ\!\!-\!\!\ast k$ and $l \ast\!\!-\!\!\circ j$ in H with $i \notin J$, and i and k are not adjacent in H, then orient $l \ast\!\!\rightarrow j$

$\mathcal{R}4$ if $(i, j, q_1, \ldots, q_n, k)$ is a discriminating path in H for j, and if $i \ast\!\!-\!\!\circ j$ in H, then orient $i \leftarrow j$ if $j \in \texttt{sepset}(\{i, k\})$ and orient $i \leftrightarrow j \leftrightarrow q_1$ if $j \notin \texttt{sepset}(\{i, k\})$

until none of these orientation rules is applicable
repeat

$\mathcal{R}5$ if $i \circ\!\!-\!\!\circ j$ in H, and there is an uncovered circle path $i \circ\!\!-\!\!\circ k \circ\!\!-\!\!\circ \cdots \circ\!\!-\!\!\circ l \circ\!\!-\!\!\circ j$ in H such that i is not adjacent to l and j is not adjacent to k, then orient $i - k - \cdots - l - j - i$

until this orientation rule is not applicable
repeat

$\mathcal{R}6$ if $i - j \circ\!\!-\!\!\ast k$ in H, then orient $j -\!\!\ast k$

$\mathcal{R}7$ if $i \ast\!\!-\!\!\circ j \circ\!\!- k$ in H with $i \notin J$, and i and k are not adjacent in H, then orient $i \ast\!\!- j$

until none of these orientation rules is applicable
repeat

$\mathcal{R}8$ if $i \rightarrow j \rightarrow k$ in H, and $i \circ\!\!\rightarrow k$ in H, then orient $i \rightarrow k$

$\mathcal{R}9$ if $i \circ\!\!\rightarrow k$, and $\pi = (i, j, \ldots, k)$ is an uncovered possibly directed path in H from i to k such that j and k are not adjacent in H, then orient $i \rightarrow k$

$\mathcal{R}10$ if $i \circ\!\!\rightarrow k$ in H, $j \rightarrow k \leftarrow l$ in H, π_1 is a uncovered possibly directed path in H from i to j, and π_2 is a uncovered possibly directed path in H from i to l, then let u_1 be the node adjacent to i on π_1 (possibly $u_1 = j$) and u_2 the node adjacent to i on π_2 (possible $u_2 = l$); if $u_1 \neq u_2$, and u_1 and u_2 are not adjacent in H, then orient $i \rightarrow k$

until none of these orientation rules is applicable

The skeleton phase in line 3, which invokes Algorithm 4, is sound (Theorem 12.7.7). That is, it computes $H = \mathrm{skel}(\mathrm{PAG}^{\sigma}(G \mid S))$, and `sepset` will contain a separating set of i and j for every edge $i \circ\!\!-\!\!\circ j$ absent in H with $i \in V, j \in J \cup V, i \neq j$. The next step, line 4, orients the edges between input and output nodes in H. The result is then a valid σ-PAG that represents G given S. The rest of the proof proceeds by induction.

Now that the skeleton has been determined, the orientations (edge marks) of the edges will be deduced. We show for each of the orientation rules that under the assumption that the current H is a valid σ-PAG that represents G given S, applying the rule yields an updated H that is still a valid σ-PAG that represents G given S. We first exploit the modeling assumptions on the input nodes in line 4 to partially orient edges connecting input and output nodes. The soundness of this orientation step stems from Lemma 12.2.8.

Some of the rules ($\mathcal{R}1$, $\mathcal{R}3$, $\mathcal{R}5$, $\mathcal{R}6$, $\mathcal{R}7$, $\mathcal{R}9$, $\mathcal{R}10$) assume that rule $\mathcal{R}0$ has been exhaustively applied, which is the reason that rule $\mathcal{R}0$ is performed before the other orientation rules are performed. Additionally, rule $\mathcal{R}6$ assumes that rule $\mathcal{R}5$ has been exhaustively applied.

In the following, we will always assume that the antecedent of the rule holds for a mixed graph H that is a valid σ-PAG that represents CDMG G given S. This implies, in particular, that if $v \leftarrow\!\!* w$ in H or $v \circ\!\!-\!\!* w$ in H, then $v \notin J$.

$\mathcal{R}0$ "If $i *\!\!-\!\!* j *\!\!-\!\!* k$ in H, with $i \notin J$, and i and k are not adjacent in H, then orient $i *\!\!\rightarrow j \leftarrow\!\!* k$ if $j \notin$ `sepset`$(\{i,k\})$."
It follows from Proposition 12.4.1 that H still represents G given S after the orientation of the unshielded collider.

$\mathcal{R}1$ "If $i *\!\!-\!\!\circ j \leftarrow\!\!* k$ in H with $i \notin J$, and i and k are not adjacent in H, then orient $i \leftarrow j$."
We first show that $j \in \mathrm{Anc}_G(i)$. Since the triple (i,j,k) is an unshielded triple in H with $i \notin J$, but has not been oriented as a collider by $\mathcal{R}0$, we conclude that $j \in$ `sepset`$(\{i,k\})$. By Proposition 12.4.1, $j \in \mathrm{Anc}_G(\{i,k\} \cup S)$. Since H represents G and $j \leftarrow\!\!* k$ in H, $j \notin \mathrm{Anc}_G(\{k\} \cup S)$. Therefore, $j \in \mathrm{Anc}_G(i)$. Thus if we orient $i *\!\!- j$, the mixed graph still represents G given S. In particular, the orientation $i - j \leftarrow\!\!* k$ with i,k non-adjacent cannot occur. Therefore, if we orient $i \leftarrow j$, the resulting mixed graph H will still represent G.

$\mathcal{R}2$ "If $i \rightarrow j *\!\!\rightarrow k$ or $i *\!\!\rightarrow j \rightarrow k$ in H, and $i *\!\!-\!\!\circ k$ in H, then orient $i *\!\!\rightarrow k$."
By the antecedent of the rule, and since H represents G, we have $k \notin \mathrm{Anc}_G(\{j\} \cup S)$. In case $i \rightarrow j *\!\!\rightarrow k$, we have $i \in \mathrm{Anc}_G(\{j\} \cup S)$, so if k were in $\mathrm{Anc}_G(i)$, it would follow that $k \in \mathrm{Anc}_G(\{j\} \cup S)$, a contradiction. In case $i *\!\!\rightarrow j \rightarrow k$, we have $j \in \mathrm{Anc}_G(\{k\} \cup S)$, so if k were in $\mathrm{Anc}_G(i)$, it would follow that $j \in \mathrm{Anc}_G(\{i\} \cup S)$, a contradiction. Hence, in both cases, we must have $k \notin \mathrm{Anc}_G(i)$. In both cases, we also have $k \notin \mathrm{Anc}_G(S)$ because of the arrowhead on $j *\!\!\rightarrow k$. Therefore, after orienting $i *\!\!\rightarrow k$ in H, the resulting mixed graph still represents G.

$\mathcal{R}3$ "If $i *\!\!\rightarrow j \leftarrow\!\!* k$ and $i *\!\!-\!\!\circ l \circ\!\!-\!\!* k$ and $l *\!\!-\!\!\circ j$ in H with $i \notin J$, and i and k are not adjacent in H, then orient $l *\!\!\rightarrow j$."

Since (i, l, k) is an unshielded triple in H with $i \notin J$ that was not oriented as a collider by $\mathcal{R}0$, we must have that $l \in \mathrm{Anc}_G(\{i, k\} \cup S)$ by Proposition 12.4.1. Assume, for the sake of contradiction, that $j \in \mathrm{Anc}_G(\{l\} \cup S)$. Then $j \in \mathrm{Anc}_G(\{i, k\} \cup S)$. This contradicts that $j \notin \mathrm{Anc}_G(\{i, k\} \cup S)$ from $i \;{*}\!\!\rightarrow\; j \;\leftarrow\!\!{*}\; k$ in H. Hence, $j \notin \mathrm{Anc}_G(l \cup S)$. After orienting $l \;{*}\!\!\rightarrow\; j$ in H, the resulting mixed graph still represents G.

$\mathcal{R}4$ "If $(i, j, q_1, \ldots, q_n, k)$ is a discriminating path in H for j, and if $i \;{*}\!\!-\!\!\circ\; j$ in H, then orient $i \leftarrow j$ if $j \in \texttt{sepset}(\{i, k\})$ and orient $i \leftrightarrow j \leftrightarrow q_1$ if $j \notin \texttt{sepset}(\{i, k\})$."
It follows immediately from Proposition 12.5.3 that applying this rule yields an updated mixed graph that still represents G.

$\mathcal{R}5$ "If $i \;\circ\!\!-\!\!\circ\; j$ in H, and there is an uncovered circle path $i \;\circ\!\!-\!\!\circ\; k \;\circ\!\!-\!\!\circ\; \cdots \;\circ\!\!-\!\!\circ\; l \;\circ\!\!-\!\!\circ\; j$ in H such that i is not adjacent to l and j is not adjacent to k, then orient $i - k - \cdots - l - j - i$."
There is an uncovered cycle consisting of (at least 4) $\circ\!\!-\!\!\circ$ edges in H. Lemma 12.8.2 implies that each node on the uncovered cycle must be in $\mathrm{Anc}_G(S)$. Hence we can orient all edges on the cycle as undirected, and this yields a mixed graph that still represents G given S.

$\mathcal{R}6$ "If $i - j \;\circ\!\!-\!\!{*}\; k$ in H, then orient $j \;-\!\!{*}\; k$."
Only $\mathcal{R}5$ could have introduced the undirected edge. In that case, we know that both i and j are in $\mathrm{Anc}_G(S)$ from Lemma 12.8.2. Hence we can orient $j \;-\!\!{*}\; k$ and the updated mixed graph will still represent G given S.

$\mathcal{R}7$ "If $i \;{*}\!\!-\!\!\circ\; j \;\circ\!\!-\; k$ in H with $i \notin J$, and i and k are not adjacent in H, then orient $i \;{*}\!\!-\; j$."
Suppose after orienting $i \;{*}\!\!\rightarrow\; j \;\circ\!\!-\; k$, the mixed graph would still represent G given S. If $j \in \mathrm{Anc}_G(\{k\} \cup S)$, then we could further orient $i \;{*}\!\!\rightarrow\; j - k$ and the mixed graph would still represent G given S, yielding a contradiction because an unshielded triple of that form cannot occur. So $j \notin \mathrm{Anc}_G(\{k\} \cup S)$, and we can orient $i \;{*}\!\!\rightarrow\; j \leftarrow k$ to obtain a mixed graph that still represents G given S. But then we have an unshielded collider (i, j, k) with $i \notin J$ that should have been oriented by $\mathcal{R}0$, another contradiction. Hence we can orient $i \;{*}\!\!-\; j$ and the resulting mixed graph will still represent G given S.

$\mathcal{R}8$ "If $i \rightarrow j \rightarrow k$ in H, and $i \;\circ\!\!\rightarrow\; k$ in H, then orient $i \rightarrow k$."
It follows immediately from Lemma 12.3.6 that $i \in \mathrm{Anc}_G(\{k\} \cup S)$. Thus, applying this rule yields an updated mixed graph that still represents G.

$\mathcal{R}9$ "If $i \;\circ\!\!\rightarrow\; k$, and $\pi = (i, j, \ldots, k)$ is an uncovered possibly directed path in H from i to k such that j and k are not adjacent in H, then orient $i \rightarrow k$."
Note that $j \notin J$ as either $i \;{*}\!\!\rightarrow\; j$ or $i \;{*}\!\!-\!\!\circ\; j$. The unshielded triple (j, i, k), with $j \notin J$, was not oriented as a collider by $\mathcal{R}0$, and therefore $i \in \mathrm{Anc}_G(\{j, k\} \cup S)$. If $i \in \mathrm{Anc}_G(j)$, then Lemma 12.8.3 then states that $i \in \mathrm{Anc}_G(k)$. Therefore, we

conclude that $i \in \mathrm{Anc}_G(\{k\} \cup S)$. Applying the rule therefore yields an updated mixed graph that still represents G.

$\mathcal{R}10$ "Suppose that $i \circ\!\!\rightarrow k$ in H, $j \rightarrow k \leftarrow l$ in H, π_1 is a uncovered possibly directed path in H from i to j, and π_2 is an uncovered possibly directed path in H from i to l. Let u_1 be the node adjacent to i on π_1 (possibly $u_1 = j$) and u_2 the node adjacent to i on π_2 (possibly $u_2 = l$). If $u_1 \neq u_2$, and u_1 and u_2 are not adjacent in H, then orient $i \rightarrow k$."
Because u_1 lies on a possibly directed path starting at i, it cannot be in J; the same holds for u_2. The unshielded triple (u_1, i, u_2), with $u_1, u_2 \notin J$, was not oriented by rule $\mathcal{R}0$, which implies that $i \in \mathrm{Anc}_G(\{u_1, u_2\} \cup S)$. If $i \in \mathrm{Anc}_G(u_1)$, Lemma 12.8.3 gives that $i \in \mathrm{Anc}_G(j)$. If $i \in \mathrm{Anc}_G(u_2)$, Lemma 12.8.3 gives that $i \in \mathrm{Anc}_G(l)$. In both cases, $i \in \mathrm{Anc}_G(\{k\} \cup S)$, because $j \rightarrow k \leftarrow l$ in H. So we conclude that $i \in \mathrm{Anc}_G(\{k\} \cup S)$ and can orient $i \rightarrow k$ to obtain an updated mixed graph that still represents G given S.

Since each of these orientation rules leaves the skeleton (adjacencies) of H invariant, and after each orientation rule, H still represents G, H remains a valid σ-PAG throughout the orientation phase by Proposition 12.3.7. $\square$

The following two lemmata are applicable after the initial phase of the extended FCI algorithm, once rule $\mathcal{R}0$ has been exhaustively applied. The first concerns uncovered circle paths, the second uncovered possibly directed paths.

Lemma 12.8.2. *Let H be a mixed graph that represents G given S in which rule $\mathcal{R}0$ has been exhaustively applied. If $i \circ\!\!-\!\!\circ j$ in H, and there is an uncovered circle path $i \circ\!\!-\!\!\circ k \circ\!\!-\!\!\circ \cdots \circ\!\!-\!\!\circ l \circ\!\!-\!\!\circ j$ in H such that i is not adjacent to l and j is not adjacent to k, then every node on the path is in $\mathrm{Anc}_G(S)$.*

Proof. There is an uncovered cycle consisting of (at least 4) $\circ\!\!-\!\!\circ$ edges in H. Note that none of the nodes on the cycle can be in J, as each node has two or more circle edge marks. Suppose we orient one of the circles into an arrowhead and the new mixed graph still represents G given S. Then we could make use of rule $\mathcal{R}1$ repeatedly to orient the whole cycle as a directed cycle, and the resulting mixed graph should still represent G given S. However, that would be a contradiction, since it contains a directed cycle. Hence, any circle edge mark on the cycle that we orient as an arrowhead would yield a mixed graph that no longer represents G given S. In other words, each node on the cycle must be ancestor in G of its neighboring node, or of the selection set S. This implies that every node on the cycle must be in $\mathrm{Anc}_G(S)$. Indeed, if a given node on the cycle is not in $\mathrm{Anc}_G(S)$, then it must be ancestor in G of its neighbors. Its neighbors then cannot be in $\mathrm{Anc}_G(S)$ either, and therefore must be ancestor in G of their neighbors. But then we would have an unshielded triple of nodes that are in the same strongly connected component of G: a contradiction with Lemma 12.2.2. $\square$

Lemma 12.8.3. *Let H be a mixed graph that represents G given S in which rule $\mathcal{R}0$ has been exhaustively applied. Let $v_1 *\!\!-\!\!* \ldots *\!\!-\!\!* v_n$ be an uncovered possibly directed* *path from v_1 to v_n (with $n \geq 2$) in H. If there is an edge $k \ast\!\!\rightarrow v_1$ in H and if $v_1 \in \mathrm{Anc}_G(\{v_2\} \cup S)$, then we conclude that $v_1 \in \mathrm{Anc}_G(v_n)$.*

Proof. Note that $v_1, v_2, \ldots, v_n$ are all not in J, because each of them must have an arrow head or circle edge mark. We can orient $v_1 \multimap\!\!\ast\, v_2$ because of the assumption that $v_1 \in \mathrm{Anc}_G(\{v_2\} \cup S)$. Note that $v_1 \in \mathrm{Anc}_G(v_2)$ because $v_1 \notin \mathrm{Anc}_G(S)$ due to the arrowhead at v_1 on the edge $k \ast\!\!\rightarrow v_1$. If $n = 2$, then we immediately conclude that $v_1 \in \mathrm{Anc}_G(v_n)$. So assume $n > 2$. We distinguish two cases: $v_2 \in \mathrm{Anc}_G(v_1)$ and $v_2 \notin \mathrm{Anc}_G(v_1)$.

If $v_2 \notin \mathrm{Anc}_G(v_1)$, we can orient $v_1 \rightarrow v_2$, since $v_2 \in \mathrm{Anc}_G(S)$ would imply $v_1 \in \mathrm{Anc}_G(S)$, a contradiction. By the same reasoning as in the proof of rule $\mathcal{R}1$, we conclude that $v_2 \in \mathrm{Anc}_G(v_3)$, and that we can orient $v_2 \rightarrow v_3$. We can iterate this reasoning subsequently on all remaining edges on the uncovered possibly directed path to deduce that $v_i \in \mathrm{Anc}_G(v_{i+1})$ and we can orient $v_i \rightarrow v_{i+1}$, for $i = 3, \ldots, n-1$. This leads to the conclusion that $v_1 \in \mathrm{Anc}_G(v_n)$.

If $v_2 \in \mathrm{Anc}_G(v_1)$, that is $v_2 \in \mathrm{Sc}_G(v_1)$, then we can orient $v_1 - v_2$. By Lemma 12.2.7, this implies that k and v_2 must be adjacent in H as well. This edge can be oriented as $k \ast\!\!\rightarrow v_2$, because $v_2 \in \mathrm{Anc}_G(\{k\} \cup S)$ would imply $v_1 \in \mathrm{Anc}_G(\{k\} \cup S)$, a contradiction. The assumption $v_2 \notin \mathrm{Anc}_G(\{v_3\} \cup S)$ gives a contradiction: we could then orient $v_2 \leftarrow\!\!\ast\, v_3$, obtaining an unshielded triple $v_1 - v_2 \leftarrow\!\!\ast\, v_3$. Hence $v_2 \in \mathrm{Anc}_G(\{v_3\} \cup S)$, and we can orient $v_2 \multimap\!\!\ast\, v_3$. Because of the arrowhead in $k \ast\!\!\rightarrow v_2$ at v_2, $v_2 \notin \mathrm{Anc}_G(S)$, and hence $v_2 \in \mathrm{Anc}_G(v_3)$. If $v_3 \in \mathrm{Anc}_G(\{v_2\} \cup S)$, then $v_3 \in \mathrm{Sc}_G(v_2)$ (as $v_3 \in \mathrm{Anc}_G(S) \setminus \mathrm{Anc}_G(v_2)$ would contradict $v_2 \notin \mathrm{Anc}_G(S)$). But then v_1 and v_3 must lie in the same strongly connected component of G, and should therefore be adjacent in H by Lemma 12.2.2, which they are not. Hence $v_3 \notin \mathrm{Anc}_G(\{v_2\} \cup S)$ and we can orient $v_2 \rightarrow v_3$. The reasoning now proceeds as in the previous case, and leads to the conclusion that $v_2 \in \mathrm{Anc}_G(v_n)$. Because $v_1 \in \mathrm{Anc}_G(v_2)$, also $v_1 \in \mathrm{Anc}_G(v_n)$. □

The soundness of the Extended FCI algorithm immediately implies its consistency when using consistent conditional independence tests.

Corollary 12.8.4 (Extended FCI Consistency)**.** *Let $M = (J, V^+, W, \mathcal{X}, P, f)$ be a simple SCM with endogenous variables $V^+ = V \dot\cup S \dot\cup L$ (note that we allow additional latent endogenous variables L here). Let $\xi_S \subseteq \mathcal{X}_S$ be a measurable set with $\mathbb{P}_M(X_S \in \xi_S \,|\, \mathrm{do}(X_J = x_J)) > 0$ for all $x_j \in \mathcal{X}_J$. Assume that we have access to infinitely many samples distributed according to the marginal Markov kernel of M on V after selecting on S:*

$$P_M(X_V \mid X_S \in \xi_S, \mathrm{do}(X_J = x_J)).$$

Assume also that the following faithfulness assumption holds:

$$\mathrm{IM}(P_M(X_V \mid X_S \in \xi_S, \mathrm{do}(X_J = x_J))) = \mathrm{IM}_\sigma(G_{V \cup S \,|\, J}(M) \,|\, S).$$

When using asymptotically consistent conditional independence tests (on the i.i.d. samples of the Markov kernel $P_M(X_V \mid X_S \in \xi_S, \mathrm{do}(X_J = x_J))$), the Extended FCI algorithm (Algorithm 5) provides an asymptotically consistent estimate $\hat{H}$ of the σ-PAG $\mathtt{FCI}(\mathrm{IM}_\sigma(G_{V \cup S \,|\, J}(M) \,|\, S))$, which represents $G_{V \cup S \,|\, J}(M)$ given S.

Proof. The asymptotic consistency of the conditional independence tests means that the probability of a wrong test result (either Type I or Type II error) vanishes asymptotically. Since $\mathrm{IM}(P_M(X_V \mid X_S \in \xi_S, \mathrm{do}(X_J = x_J)))$ consists of finitely many conditional independence statements, the test results will agree completely with this conditional independence model with arbitrarily high probability given sufficiently many samples. By the faithfulness assumption, this conditional independence model agrees with $\mathrm{IM}_\sigma(G_{V \cup S \,|\, J}(M) \,|\, S)$. By the soundness of the Extended FCI algorithm (Theorem 12.8.1), the σ-PAG that FCI outputs will equal $\texttt{FCI}(\mathrm{IM}_\sigma(G_{V \cup S \,|\, J}(M) \,|\, S))$ with arbitrarily high probability given sufficiently many samples. □

Because conditional independence tests are not uniformly consistent (as there is no upper bound on the number of samples needed to distinguish an arbitrarily weak dependence from an independence, without additional assumptions), also the Extended FCI algorithm is not uniformly consistent. In other words, it is not known in advance how many samples will be needed to yield a reliable result.

12.9. Completeness

For acyclic G without input nodes (that is, if G is an ADMG), the FCI algorithm was shown to be complete [Zha08] in the sense that all edge marks that could possibly be oriented based on the information in $\mathrm{IM}_\sigma(G \,|\, S)$ will be oriented. Using results on the characterization of Markov equivalence classes of maximal ancestral graphs [ARSZ05], it can additionally be shown that the σ-PAG output by FCI represents the Markov equivalence class of G with respect to V in case G is an ADMG. By employing acyclifications, these results have been extended to cyclic G [MC20], still without input nodes, and with the additional assumption of no selection bias ($S = \emptyset$). The known completeness results have very long proofs, and we will therefore not provide these here. Instead, we will only formulate these results, and refer the interested reader to the original papers for the proofs.

We make the following assumption in order to state the known completeness results.[87]

Assumption 12.9.1. *Given node set V, let:*

- *$\mathcal{G}_V$ be the set of all pairs (G, S) of acyclic DMGs G with output nodes $V^+ = V \,\dot{\cup}\, S$ for some disjoint set S ('acyclicity'), or*
- *$\mathcal{G}_V$ be the set of all pairs $(G, \emptyset)$ of DMGs G with output nodes $V^+ = V \,\dot{\cup}\, \emptyset$ ('no selection bias').*

In both cases, all DMGs in $\mathcal{G}_V$ have $J = \emptyset$ ('no input nodes').

[87] While we believe that it is straightforward to extend the completeness results to allow for both cycles and selection bias, it is known that the Extended FCI algorithm (Algorithm 5) is incomplete when also allowing for input nodes.

In the absence of input nodes (for $J = \emptyset$), we can interpret FCI as a mapping that maps an independence model (on V) to a mixed graph (with nodes V). By Theorem 12.8.1, it maps $\mathrm{IM}_\sigma(G \mid S)$, the σ-independence model of a DMG G given S, to a σ-PAG $\texttt{FCI}(\mathrm{IM}_\sigma(G \mid S))$ that represents G given S. Additionally, the following (incomplete) completeness results are known.

Theorem 12.9.2 (Some FCI completeness results)**.** *Under Assumption 12.9.1, the Extended FCI algorithm (Algorithm 5) is:*

(i) ***arrowhead complete****: for all $(G, S) \in \mathcal{G}_V$, for all $i \neq j \in V$: there is an arrowhead $i \leftarrow\!\!* \; j$ in $\texttt{FCI}(\mathrm{IM}_\sigma(G \mid S))$ if $i \notin \mathrm{Anc}_{\tilde{G}}(\{j\} \cup S)$ for all $(\tilde{G}, \tilde{S}) \in \mathcal{G}_V$ with $\mathrm{IM}_\sigma(\tilde{G} \mid \tilde{S}) = \mathrm{IM}_\sigma(G \mid S)$;*

(ii) ***tail complete****: for all $(G, S) \in \mathcal{G}_V$, for all $i \neq j \in V$: there is a tail $i \;-\!\!* \; j$ in $\texttt{FCI}(\mathrm{IM}_\sigma(G \mid S))$ if $i \in \mathrm{Anc}_{\tilde{G}}(\{j\} \cup S)$ for all $(\tilde{G}, \tilde{S}) \in \mathcal{G}_V$ with $\mathrm{IM}_\sigma(\tilde{G} \mid \tilde{S}) = \mathrm{IM}_\sigma(G \mid S)$;*

(iii) ***Markov complete****: for all $(G_1, S_1) \in \mathcal{G}_V$ and $(G_2, S_2) \in \mathcal{G}_V$: $\mathrm{IM}_\sigma(G_1 \mid S_1) = \mathrm{IM}_\sigma(G_2 \mid S_2)$ if and only if $\texttt{FCI}(\mathrm{IM}_\sigma(G_1 \mid S_1)) = \texttt{FCI}(\mathrm{IM}_\sigma(G_2 \mid S_2))$.*

Proof. The first two claims are proved in [Zha08] under the additional assumption of acyclicity. In [MC20] it is explained how the characterization of the Markov equivalence classes of [ARSZ05] can be used to then prove the third claim under that additional assumption. Furthermore, in [MC20] it is shown how to generalize these results to the cyclic setting by employing acyclifications, but only under the additional assumption of no selection bias. □

Arrowhead and tail completeness express that the σ-PAG output by FCI is maximally oriented: any arrowhead or tail that could possibly be deduced from $\mathrm{IM}_\sigma(G \mid S)$, will have been oriented as such in the σ-PAG. The soundness and Markov completeness properties together imply that the σ-PAG output by FCI, when given as input the σ-independence model of a directed mixed graph given some set of latent selection nodes, represents the σ-Markov equivalence class of G with respect to the observed nodes. In other words, FCI provides a *graphical characterization* of the σ-Markov equivalence class.

A. Appendix: Measure Theoretic Probability

This appendix provides a crash course (or rather, refresher) of concepts from measure theoretic probability.

A.1. Why Measure Theory?

Discrete and absolute continuous distributions are not general enough

Example A.1.1 (Simple example of a non-discrete non-absolute-continuous distribution)**.** *Consider a uniformly distributed random variable on the interval $\mathcal{X} := [0, 1]$, i.e. $X \sim \mathcal{U}[0, 1]$, which has probability density:*

$$p(x) = \mathbb{1}_{[0,1]}(x).$$

Consider an exact copy of X, which we call $Y := X$, on $\mathcal{Y} := [0, 1]$. Now consider the joint distribution of (X, Y) on $\mathcal{X} \times \mathcal{Y} = [0, 1]^2$. Then only values on the diagonal $\Delta := \{(x, x) \mid x \in [0, 1]\}$ can be realized by (X, Y). This simple distribution on $[0, 1]^2$ is not discrete (as it can attain uncountably many values), and it is also not absolute continuous, since we have: $\int_\Delta dx\, dy = 0$, i.e. the (2-dimensional) area of the (1-dimensional) line is zero. This implies that any density function p would satisfy: $\int_\Delta p(x, y)\, dx\, dy = 0$ as well. This is in contrast to the fact that a probability distribution should always be normalized:

$$1 = P((X, Y) \in \Delta) = \int_\Delta p(x, y)\, dx\, dy.$$

Note that we don't need a probability density to be able to assign probabilities to subsets $D \subseteq [0, 1]^2$. We can just use the push-forward map:

$$(X, Y) : \ [0, 1] \to [0, 1] \times [0, 1], \quad x \mapsto (x, x).$$

and compute:

$$P((X, Y) \in D) = P(\{x \in [0, 1] \mid (x, x) \in D\}),$$

where P on the right here denotes the uniform distribution on $[0, 1]$.

Notation A.1.2 (Unifying the notations to measure theoretic ones)**.** *Let X be a random variable taking values in space $\mathcal{X}$ and with probability distribution P. Let $F : \mathcal{X} \to \mathbb{R}$ be a function. Then we will change the notations for expectation values as follows.*

1. *Let X be a **discrete** random variable with **probability mass function** p. Then define:*

$$\begin{aligned}\mathbb{E}[F(X)] &= \sum_{x \in \mathcal{X}} F(x) \cdot p(x) \\ &=: \int F(x)\, P(dx) \\ &=: \int F(x)\, dP(x) \\ &=: \int F\, dP.\end{aligned}$$

We will consider sums to be special cases of measure integrals.

2. *Let X be a **absolute continuous** random variable with **probability density function** p. Then define:*

$$\begin{aligned}\mathbb{E}[F(X)] &= \int_{\mathcal{X}} F(x) \cdot p(x)\, dx \\ &=: \int F(x)\, P(dx) \\ &=: \int F(x)\, dP(x) \\ &=: \int F\, dP.\end{aligned}$$

So both cases can be unified with the 3 commonly used notations:

$$\mathbb{E}[F(X)] = \int F\, dP = \int F(x)\, dP(x) = \int F(x)\, P(dx).$$

Note that in both cases we also can write: $P(A) = \int \mathbb{1}_A\, dP$.

Exercise A.1.3. *Show that the following relation holds:*

$$\int F(x)\, P(dx) = \int z\, P^F(dz).$$

Defining probability distributions on all subsets is too general

Remark A.1.4. *When we want to work with a (probability) measure μ we at least want to require that it is **countably additive**, i.e. that for pairwise **disjoint** subsets $A_n \subseteq \mathcal{X}$, $n \in \mathbb{N}$, we have:*

$$\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

We will see below that if we do not restrict the subsets A_n in some way we will encounter strange behaviour.

Theorem A.1.5 (Vitali, non-existence of Lebesgue measure on all subsets)**.** *There does NOT exist a measure λ on $[0,1]$ such that:*

1. *λ can measure every subset $A \subseteq [0,1]$, and:*

2. *$\lambda([a,b]) = b - a$ for all $a \leq b$ with $a, b \in [0,1]$.*

In other words, there does NOT exist a uniform distribution on $[0,1]$ that can consistently assign values to all subsets.
*But: such a measure with property 2. exists on the set $\mathcal{B}_{[0,1]}$ of all so called **Borel subsets** (or even Lebesgue subsets) of $[0,1]$. Similar statements hold for higher dimensions and $\mathbb{R}^D$ and higher dimensional volumes.*

Example A.1.6 (Vitali set). *Consider the following equivalence relation on* $[0,1]$*:*

$$r_1 \sim r_2 \qquad :\Longleftrightarrow \qquad r_2 - r_1 \in \mathbb{Q}.$$

Let $[0,1]/\sim$ *be the set of equivalence classes. By the axiom of choice there exists a representative system* $V \subseteq [0,1]$ *for* $[0,1]/\sim$*. This means that the map:*

$$V \to [0,1]/\sim, \qquad v \mapsto [v],$$

is bijective. V *is called* ***Vitali set*** *and we claim that* V *is not Lebesgue-measurable. For this let* $q \in \mathbb{Q}$ *and consider the subset:*

$$V_q := V + q := \{v + q \,|\, v \in V\} \subseteq \mathbb{R}.$$

Let $\mathcal{Q} := [-1,1] \cap \mathbb{Q}$*. Note that* $[0,1]$ *and* V_q *are uncountable while* $\mathbb{Q}$ *and* $\mathcal{Q}$ *are countably infinite. We then have the inclusions:*

$$[0,1] \subseteq \bigcup_{q \in \mathcal{Q}} V_q \subseteq [-1,2].$$

The right inclusion is clear as:

$$V + [-1,1] \subseteq [0,1] + [-1,1] \subseteq [-1,2].$$

For the left inclusion let $x \in [0,1]$*. By construction there exists a* $v \in V$ *such that* $v \sim x$*. So* $x - v \in \mathbb{Q}$*. Since* $x, v \in [0,1]$ *we also have that* $x - v \in [-1,1]$*. So* $q := x - v \in [-1,1] \cap \mathbb{Q} = \mathcal{Q}$*. This shows that* $x \in V_q$ *for a* $q \in \mathcal{Q}$*. Thus both inclusions are shown.*
If we now assumed that V *would be Lebesgue-measurable then every* V_q *would be as well as a translated version of* V*. We then would get that:* $\lambda(V_q) = \lambda(V)$ *for every* $q \in \mathbb{Q}$*. So we would get:*

$$1 = \lambda([0,1]) \leq \lambda\left(\bigcup_{q \in \mathcal{Q}} V_q\right) \leq \lambda([-1,2]) = 3,$$

which implies:

$$[1,3] \ni \lambda\left(\bigcup_{q \in \mathcal{Q}} V_q\right) = \sum_{q \in \mathcal{Q}} \lambda(V_q) = \sum_{q \in \mathcal{Q}} \lambda(V),$$

which is contradictory. Indeed, $\lambda(V) = 0$ *can be ruled out as the sum would sum up to* $0 \notin [1,3]$*. But also* $\lambda(V) > 0$ *can be ruled out as this would sum up to* $\infty \notin [1,3]$*. So the* ***Vitali set*** V *can not be Lebesgue-measurable.*

Theorem A.1.7 (Banach-Tarski paradox). *The 3-dimensional unit ball* $B_1(z) = \{x \in \mathbb{R}^3 \,|\, \|x - z\| \leq 1\}$ *centered at* $z \in \mathbb{R}^3$ *can be partitioned into a finite number of disjoint sets* $A_1, \dots, A_K$ *(e.g.* $K = 5$*) such that each can then be rotated and translated in* $\mathbb{R}^3$ *such that they form TWO 3-dimensional unit balls* $B_1(y_1)$ *and* $B_1(y_2)$*.*
Note that the unit balls have well-defined volume (i.e. 3-dimensional Lebesgue measure) and translation and rotations are very well behaved and preserve volume, while the subsets A_k *are very pathological (i.e. non-Lebesgue-measurable).*

Figure 34: Illustration of the Banach-Tarski paradox.[88]

$\Longrightarrow$ **Measure theory is the unifying 'safe space' for probability theory!**

A.2. Core Concepts

Motivation A.2.1. *As discussed before in remark A.1.4, we want to define probability measures P on a space $\mathcal{W}$. We want them to follow (at least) these rules:*

i) normalized: $P(\mathcal{W}) = 1$, $P(\emptyset) = 0$.

ii) complement: $P(A^\complement) = 1 - P(A)$ *for* $A \subseteq \mathcal{W}$.

iii) σ-additivity (aka countably additivity): For pairwise disjoint subsets $A_n \subseteq \mathcal{W}$, $n \in \mathbb{N}$*:*

$$P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n).$$

Such rules implicitly assume that P can measure the sets $\mathcal{W}$ and $\emptyset$; and that P can measure the complement $A^\complement$ if it can measure A; and that P can measure the (disjoint) union $\bigcup_{n\in\mathbb{N}} A_n$ if it can measure each of the A_n.
As illustrated by the theorems A.1.5 and A.1.7, this is in general NOT possible to do for all subsets of $\mathcal{W}$ (i.e. for all elements of the power set $2^{\mathcal{W}}$).
*This problem is solved and formalized by the notion of **σ-algebras** of subsets of the space $\mathcal{W}$.*

Definition A.2.2 (σ-algebras)**.** *Let $\mathcal{W}$ be a set. A (non-empty) set $\mathcal{B} \subseteq 2^{\mathcal{W}}$ of subsets $A \subseteq \mathcal{W}$ is called a **σ-algebra** on $\mathcal{W}$ if it satisfies the following rules:*

i) empty set: $\emptyset \in \mathcal{B}$,

ii) complement: If $A \in \mathcal{B}$ *then also:* $A^\complement := \mathcal{W} \setminus A \in \mathcal{B}$,

iii) countable union: If $A_n \in \mathcal{B}$ *for all* $n \in \mathbb{N}$ *then also:* $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{B}$.

Definition A.2.3 (Measurable spaces)**.** *A tuple $(\mathcal{W}, \mathcal{B})$ of a set $\mathcal{W}$ and a σ-algebra $\mathcal{B}$ on $\mathcal{W}$ is called **measurable space**.*

[88] https://en.wikipedia.org/wiki/Banach-Tarski_paradox

Remark A.2.4 (Abuse of notation). *By abuse of notation we often just call $\mathcal{W}$ a measurable space by implicitly assuming that it is endowed with a fixed σ-algebra, which we will indicate by $\mathcal{B}_{\mathcal{W}}$ or $\mathcal{B}(\mathcal{W})$ if needed. We will also just call a subsets $A \subseteq \mathcal{W}$* ***measurable*** *when we actually mean that $A \in \mathcal{B}_{\mathcal{W}}$.*

Definition A.2.5 (Measures). *Let $(\mathcal{W}, \mathcal{B})$ be a measurable space. A* ***measure*** *μ on $(\mathcal{W}, \mathcal{B})$ - by definition - is a mapping:*

$$\mu : \mathcal{B} \to \mathbb{R} \cup \{\infty\}, \quad D \mapsto \mu(D),$$

such that:

i) non-negative: $\forall A \in \mathcal{B}$: $\mu(A) \in [0, \infty]$,

ii) empty set: $\mu(\emptyset) = 0$,

iii) countably additive (aka σ-additive): for all sequences $A_n \in \mathcal{B}$, $n \in \mathbb{N}$, with $A_i \cap A_j = \emptyset$ for all $i \neq j$, we have:

$$\mu\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} \mu(A_n).$$

Definition A.2.6 (Probability/finite/σ-finite measures). *A measure μ on $(\mathcal{W}, \mathcal{B})$ is called:*

1. ***probability measure*** *if $\mu(\mathcal{W}) = 1$.*
2. ***finite measure*** *if $\mu(\mathcal{W}) < \infty$.*
3. ***σ-finite measure*** *if there are $D_n \in \mathcal{B}$, $n \in \mathbb{N}$, with $\mu(D_n) < \infty$ and $\mathcal{W} = \bigcup_{n \in \mathbb{N}} D_n$.*

Definition A.2.7 (Measure spaces/probability spaces). *A triple $(\mathcal{W}, \mathcal{B}, \mu)$ consisting of a measurable space $(\mathcal{W}, \mathcal{B})$ and a measure μ on $(\mathcal{W}, \mathcal{B})$ is called* ***measure space*** *(and* ***probability space*** *if μ is a probability measure).*
Again, by abuse of notation, we often omit the σ-algebra in the notation and call $(\mathcal{W}, \mu)$ a measure space, probability space, resp.

Definition A.2.8 (Measurable mappings). *Let $(\mathcal{W}, \mathcal{B}_{\mathcal{W}})$ and $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$ be two measurable spaces and $f : \mathcal{W} \to \mathcal{Z}$ be a mapping. We call f a* ***$\mathcal{B}_{\mathcal{W}}$-$\mathcal{B}_{\mathcal{Z}}$-measurable mapping*** *(or just* ***measurable*** *for short) if for all $B \in \mathcal{B}_{\mathcal{Z}}$ the pre-image $f^{-1}(B)$ is an element of $\mathcal{B}_{\mathcal{W}}$. In formulas:*

$$\forall B \in \mathcal{B}_{\mathcal{Z}} : \ f^{-1}(B) \in \mathcal{B}_{\mathcal{W}}.$$

Remember the definition of pre-image: $f^{-1}(B) := \{w \in \mathcal{W} \,|\, f(w) \in B\}$.

Definition A.2.9 (Push-forward measure)**.** *Let $X : (\mathcal{W}, \mathcal{B}_\mathcal{W}) \to (\mathcal{X}, \mathcal{B}_\mathcal{X})$ be measurable and μ a measure on $(\mathcal{W}, \mathcal{B}_\mathcal{W})$. Then we define the **push-forward measure** (aka **image measure**) of μ via:*

$$(X_*\mu)(A) := \mu^X(A) := \mu_X(A) := \mu(X)(A) := \mu(X \in A) := \mu(X^{-1}(A))$$

*for all $A \in \mathcal{B}_\mathcal{X}$. If μ is a probability distribution then the push-forward measure $\mu(X)$ is also called the (distributional) **law of** X.*

Definition A.2.10 (Random variables)**.** *A measurable mapping:*

$$X : (\mathcal{W}, \mathcal{B}_\mathcal{W}, P) \to (\mathcal{X}, \mathcal{B}_\mathcal{X})$$

*that starts from a probability space is also called **random variable**.*
The main point is that the map X comes with its own distribution P^X. We often just say: "Let X be a random variable with distribution $P^X = \cdots$", where P^X is then specified, e.g. to be a Gaussian or a categorical distribution, etc.

Definition A.2.11 (Null sets)**.** *Let $(\mathcal{W}, \mathcal{B}, \mu)$ be a measure space. A subset $M \subseteq \mathcal{W}$ is called μ-**null** or μ-**zero set** if there exists a set $N \in \mathcal{B}$ with $M \subseteq N$ and $\mu(N) = 0$.*

Definition A.2.12 (Almost surely/almost all)**.** *Let $(\mathcal{X}, \mathcal{B}, \mu)$ be a measure space and $f, g : \mathcal{X} \to \mathcal{Z}$ a measurable map. We write $f =_\mu g$ or say $f = g$ μ-**almost-surely** (a.s.) or $f(x) = g(x)$ for μ-**almost-all** $x \in \mathcal{X}$ if:*

$$\{x \in \mathcal{X} \mid f(x) \neq g(x)\} \quad \text{is a } \mu\text{-null set.}$$

Similarly, for $f \leq_\mu g$, etc..
*More generally, we say that a condition C about points $x \in \mathcal{X}$ holds μ-**almost-surely** or for μ-**almost-all** $x \in \mathcal{X}$ if the set of points where the condition does not hold is μ-null, i.e.:*

$$\{x \in \mathcal{X} \mid \neg C(x)\} \quad \text{is a } \mu\text{-null set.}$$

A.3. Default Choices for Sigma-Algebras

In this subsection we want to highlight what kind of default σ-algebras we will assume on different types of spaces and on spaces constructed from others.

Remark A.3.1 (Discrete spaces)**.** *If $\mathcal{W}$ is countable (i.e. either finite or countably infinite, e.g. like $\mathbb{Z}$, $\mathbb{Q}$ or $\mathbb{N}$ or $\{1, \dots, N\}$) then we will always implicitly assume that $\mathcal{W}$ is endowed with the power set σ-algebra: $\mathcal{B}_\mathcal{W} = 2^\mathcal{W}$ (unless stated otherwise).*

Definition A.3.2 (σ-algebra generated by a set of subsets)**.** *Let $\mathcal{W}$ be a set and $\mathcal{A} \subseteq 2^\mathcal{W}$ be any non-empty set of subsets of $\mathcal{W}$. Then we can define the σ-**algebra generated by** $\mathcal{A}$:*

$$\sigma(\mathcal{A}) := \bigcap_{\substack{\mathcal{A} \subseteq \mathcal{B} \\ \mathcal{B}\ \sigma\text{-algebra on } \mathcal{W}}} \mathcal{B},$$

as the intersection of all σ-algebras $\mathcal{B}$ on $\mathcal{W}$ that contain $\mathcal{A}$. Note that the set $\sigma(\mathcal{A})$ really is a well-defined σ-algebra on $\mathcal{W}$. $\sigma(\mathcal{A})$ is thus - by definition - the smallest σ-algebra on $\mathcal{W}$ that contains $\mathcal{A}$.

Definition A.3.3 (Borel σ-algebra on topological spaces)**.** *Let $(\mathcal{W}, \mathcal{O})$ be a topological space with set of open subsets $\mathcal{O}$ then the **Borel σ-algebra** of $(\mathcal{W}, \mathcal{O})$ is defined as the smallest σ-algebra that contains all open (and thus also all closed) subsets:*

$$\mathcal{B}_{(\mathcal{W},\mathcal{O})} := \sigma(\mathcal{O}).$$

We will always implicitly assume that every topological space is endowed with its Borel σ-algebra (unless stated otherwise).

Remark A.3.4. *Caution: Other choices of σ-algebras for topological spaces used in the literature are the Baire σ-algebra, which is generated by the zero sets of all continuous functions, or the σ-algebra generated only by its closed (countably) compact sets, or the σ-algebra of all (Radon-)universally measurable subsets.*

Lemma A.3.5 (Borel σ-algebra on $\mathbb{R}^D$)**.** *The Borel σ-algebra of $\mathbb{R}^D$ is generated by the cubes:*

$$\mathcal{B}_{\mathbb{R}^D} = \sigma\left(\{[a_1, b_1] \times \cdots \times [a_D, b_D] \mid a_d, b_d \in \mathbb{Q}, a_d \leq b_d, d = 1, \dots, D\}\right).$$

Definition/Lemma A.3.6 (σ-algebras induced by mappings)**.** *Let $f : \mathcal{W} \to \mathcal{Z}$ be any mapping.*

1. *Let $\mathcal{B}_\mathcal{Z}$ be a σ-algebra on $\mathcal{Z}$. Then the **pull-back σ-algebra** defined via:*

$$f^*\mathcal{B}_\mathcal{Z} := \left\{f^{-1}(C) \,|\, C \in \mathcal{B}_\mathcal{Z}\right\}$$

   is the smallest σ-algebra $\mathcal{B}_\mathcal{W}$ that makes f $\mathcal{B}_\mathcal{W}$-$\mathcal{B}_\mathcal{Z}$-measurable.

2. *Let $\mathcal{B}_\mathcal{W}$ be a σ-algebra on $\mathcal{W}$. Then the **push-forward σ-algebra** defined via:*

$$f_*\mathcal{B}_\mathcal{W} := \left\{C \subseteq \mathcal{Z} \,|\, f^{-1}(C) \in \mathcal{B}_\mathcal{W}\right\}$$

   is the biggest σ-algebra $\mathcal{B}_\mathcal{Z}$ that makes f $\mathcal{B}_\mathcal{W}$-$\mathcal{B}_\mathcal{Z}$-measurable.

Definition A.3.7 (Product σ-algebra)**.** *Let $(\mathcal{X}_i, \mathcal{B}_i)$ be measurable spaces, $i \in I$. Then the product space $\prod_{i \in I} \mathcal{X}_i$ is endowed with the smallest σ-algebra such that for every $j \in I$ the projection map:*

$$\mathrm{pr}_j : \prod_{i \in I} \mathcal{X}_i \to \mathcal{X}_j, \qquad (x_i)_{i \in I} \mapsto x_j,$$

is measurable. We use the symbols $\bigotimes_{i \in I} \mathcal{B}_i$ for this product σ-algebra. In symbols:

$$\bigotimes_{i \in I} \mathcal{B}_i := \sigma\left(\bigcup_{i \in I} \mathrm{pr}_i^* \mathcal{B}_i\right).$$

We will always implicitly assume that every product space is endowed with this product σ-algebra (unless stated otherwise).

Definition A.3.8 (Subspace σ-algebra)**.** *Let $(\mathcal{W}, \mathcal{B})$ be a measurable space and $\mathcal{Z} \subseteq \mathcal{W}$ be a subset. Then the* ***subspace σ-algebra*** *$\mathcal{B}_{|\mathcal{Z}}$ on $\mathcal{Z}$ is the smallest σ-algebra that makes the inclusion map $\mathcal{Z} \to \mathcal{W}$ measurable. More concretely:*

$$\mathcal{B}_{|\mathcal{Z}} := \{B \cap \mathcal{Z} \mid B \in \mathcal{B}\}.$$

We will always assume that subsets are endowed with the subspace σ-algebra (unless it is ambiguous or stated otherwise).

Definition A.3.9 (Disjoint union σ-algebra)**.** *Let $(\mathcal{X}_i, \mathcal{B}_i)$ be measurable spaces, $i \in I$, considered to be pairwise disjoint. Then the* ***disjoint union σ-algebra*** *on the disjoint union $\coprod_{i \in I} \mathcal{X}_i$ is the biggest σ-algebra $\mathcal{B}_\sqcup$ such that all inclusion maps $\mathcal{X}_i \to \coprod_{i \in I} \mathcal{X}_i$ are measurable. In symbols:*

$$\mathcal{B}_\sqcup := \left\{ E \subseteq \coprod_{i \in I} \mathcal{X}_i \,\middle|\, \forall i \in I :\; E \cap \mathcal{X}_i \in \mathcal{B}_i \right\}.$$

Definition A.3.10 (σ-algebra on the space of all probability measures)**.** *Let $(\mathcal{W}, \mathcal{B}_\mathcal{W})$ be a measurable space. We denote the space of all probability measures on $(\mathcal{W}, \mathcal{B}_\mathcal{W})$ by:*

$$\mathcal{P}(\mathcal{W}) := \{P \mid P \text{ is probability measure on } (\mathcal{W}, \mathcal{B}_\mathcal{W})\}.$$

We endow $\mathcal{P}(\mathcal{W})$ with the smallest σ-algebra $\mathcal{B}_{\mathcal{P}(\mathcal{W})}$ such that all evaluation maps:

$$\mathrm{ev}_D :\; \mathcal{P}(\mathcal{W}) \to [0,1], \qquad P \mapsto P(D)$$

are measurable for $D \in \mathcal{B}_\mathcal{W}$. In symbols:

$$\mathcal{B}_{\mathcal{P}(\mathcal{W})} := \sigma\left(\bigcup_{D \in \mathcal{B}_\mathcal{W}} \mathrm{ev}_D^* \mathcal{B}_{[0,1]} \right).$$

We will always assume that the space of probability measures $\mathcal{P}(\mathcal{W})$ is endowed with this σ-algebra (unless stated otherwise).

A.4. Standard Measurable Spaces

Definition A.4.1 (Standard measurable space)**.** *A measurable space $(\mathcal{W}, \mathcal{B})$ is called* ***standard measurable space*** *(aka standard Borel space) if it is measurably isomorphic to either:*

1. *a finite measurable space $\{1, \dots, M\}$ for some $M \in \mathbb{N}$ endowed with the power set σ-algebra $2^{\{1,\dots,M\}}$, or:*
2. *the countably infinite space $\mathbb{N}$ endowed with the power set σ-algebra $2^\mathbb{N}$, or:*
3. *the unit interval $[0,1]$ endowed with its Borel σ-algebra:*

$$\mathcal{B}_{[0,1]} = \sigma\left(\{[a,b] \mid a, b \in [0,1] \cap \mathbb{Q}, a \leq b\}\right).$$

'Measurably isomorphic' means that there is a measurable mapping that has a measurable inverse.

Theorem A.4.2 (Kuratowski et al.). *1. Every Borel subset of any complete metric space that has a countable dense subset is a standard measurable space in its Borel σ-algebra (e.g. $\mathbb{Q}^D$ is countable and dense in $\mathbb{R}^D$).*

2. *Two standard measurable spaces $\mathcal{X}$ and $\mathcal{Y}$ are measurably isomorphic iff their cardinalities $|\mathcal{X}|$, $|\mathcal{Y}|$ are equal (e.g. $\mathbb{R}^D \cong [0,1]$).*

3. *Countable disjoint unions and countable direct products of standard measurable spaces are standard measurable spaces.*

4. *If $\mathcal{W}$ is standard measurable space then the space of its probablity measures $\mathcal{P}(\mathcal{W})$ is also a standard measurable space.*

Example A.4.3. *Examples of standard measurable spaces are: $\mathbb{R}^D$, $\mathbb{Q}$, $\mathbb{Z}$, $\mathbb{N}$, $\{1,\dots,M\}$, $[0,1]$, topological manifolds, countable CW-complexes, every Borel set of any separable complete metric space.*

A.5. Measure Integrals

The construction of the measure integral $\int f\,d\mu$ follows in several steps.

Construction A.5.1 (Measure integral). *Let $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \mu)$ be a measure space.*

1. *Indicator functions: For $A \in \mathcal{B}_\mathcal{X}$ put:*

$$\int \mathbb{1}_A \, d\mu := \mu(A).$$

2. *Simple functions: For a simple function $g: \mathcal{X} \to \mathbb{R}$ given by:*

$$g(x) = \sum_{n=1}^{N} a_n \cdot \mathbb{1}_{A_n}(x),$$

where $A_n \in \mathcal{B}_\mathcal{X}$ and $a_n \in \mathbb{R}$, $n = 1, \dots, N$, we define:

$$\int g \, d\mu := \sum_{n=1}^{N} a_n \cdot \mu(A_n).$$

3. *Non-negative measurable functions: Let $h: \mathcal{X} \to [0,\infty]$ be a non-negative measurable function then we define:*

$$\int h \, d\mu := \sup_{0 \le g \le h} \int g \, d\mu \qquad \in [0,\infty],$$

where the supremum is running over all non-negative simple functions g that are smaller or equal to h.

4. *Measurable functions with well-defined integral: Let* $f : \mathcal{X} \to \bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ *be a measurable function. We then can write* $f = f_+ - f_-$ *with:*

$$f_+ := \max(f, 0) \geq 0, \qquad f_- := \max(-f, 0) \geq 0.$$

If at least one of $\int f_+ \, d\mu$, $\int f_- \, d\mu$ *is finite (i.e.* $< \infty$*) we can then define:*

$$\int f \, d\mu := \int f_+ \, d\mu - \int f_- \, d\mu \qquad \in [-\infty, \infty].$$

The only case where we cannot properly define the integral is for measurable functions $f : \mathcal{X} \to \bar{\mathbb{R}}$ *where both integrals:* $\int f_+ \, d\mu = \infty$ *and* $\int f_- \, d\mu = \infty$ *are infinite, because of the "*$\infty - \infty =?$*" problem.*

Remark A.5.2 (Riemann integral vs. measure integral)**.** *The construction of the Riemann integral (RI) and the measure integral (MI) differ only in a few points:*

1. *RI uses (infinitesimal) interval length on x-axis, while MI uses the measure content (which is also the interval length in case of the Lebesgue measure).*
2. *RI decomposes the x-axis, while MI decomposes the y-axis.*
3. *RI uses limits for the integration boundaries to integrate to infinity (if convergent), while MI takes difference of integrals (to infinity) of* f_+ *and* f_- *(if difference well-defined).*
4. *RI integrates in direction* a *to* b*, while MI integrates interval* $[a, b]$ *in an undirected fashion.*
5. *If a function is Riemann integrable (RI) (e.g. continuous) on interval* $[a, b]$ *then it is also Lebesgue integrable (MI) with the same integral value.*

Definition A.5.3 (Integrable functions)**.** *Let* $(\mathcal{X}, \mathcal{B}, \mu)$ *be a measure space. A measurable function* $f : \mathcal{X} \to \mathbb{R}$ *is called* μ***-integrable*** *if:*

$$\int |f| \, d\mu < \infty.$$

Theorem A.5.4 (Properties of the integral)**.** *Let* $(\mathcal{X}, \mathcal{B}, \mu)$ *be a measure space and* $f, g : \mathcal{X} \to \mathbb{R}$ μ*-integrable measurable functions.*

1. *For* $A \in \mathcal{B}$ *we have:* $\int \mathbb{1}_A \, d\mu = \mu(A)$.
2. *Linearity: If* $a, b \in \mathbb{R}$ *then* $a \cdot f + b \cdot g$ *is also* μ*-integrable and we have:*

$$\int (a \cdot f + b \cdot g) \, d\mu = a \cdot \int f \, d\mu + b \cdot \int g \, d\mu.$$

3. *Triangle inequality:*
$$\left| \int f \, d\mu \right| \leq \int |f| \, d\mu < \infty.$$

4. *If $f \geq_\mu 0$ then: $\int f \, d\mu \geq 0$, with equality iff $f =_\mu 0$.*

5. *Monotonicity: If $f \geq_\mu g$ then: $\int f \, d\mu \geq \int g \, d\mu$, with equality iff $f =_\mu g$.*

6. *If $\int f \, d\mu < \infty$ then $f <_\mu \infty$.*

7. *The measure integral satisfies monotone convergence, dominated convergence, Fubini theorems, etc. (see literature).*

Note, we use $=_\mu$ and $\geq_\mu$ to indicate that this property is (only) allowed to fail on a μ-null set.

Definition A.5.5 (Expectation value)**.** *Let $(\mathcal{W}, \mathcal{B}, P)$ be a probability space and $X : \mathcal{W} \to \mathbb{R}$ be a measurable function with well-defined integral. Then its* ***expectation value*** *(w.r.t. P) is defined to be:*

$$\mathbb{E}[X] := \int X \, dP.$$

Example A.5.6. *Let $\mathcal{X}$ be a measurable space and $f : \mathcal{X} \to \mathbb{R}$ a measurable function and $\mathcal{W} \subseteq \mathcal{X}$ a countable subset.*

1. ***Dirac measure****. Let $w \in \mathcal{X}$ be a point. We define the* ***Dirac measure*** *δ_w centered at w via:*
$$\delta_w(A) := \mathbb{1}_A(w),$$
for all measurable $A \subseteq \mathcal{X}$. Furthermore, we have:
$$\mathbb{E}[f] = \int f(x) \, \delta_w(dx) = f(w).$$
This holds because: $f(x) = f(w)$ for δ_w-almost-all $x \in \mathcal{X}$. Let's prove the right equality more formally:

   Proof. Consider:
$$B := f^{-1}(f(w)) = \{x \in \mathcal{X} \mid f(x) = f(w)\} \ni w.$$
Since f is measurable and $\{f(w)\} \in \mathcal{B}_\mathbb{R}$ we also have $B \in \mathcal{B}_\mathcal{X}$. We then have the decomposition:
$$\begin{aligned} f(x) &= f(x) \cdot \mathbb{1}_B(x) + f(x) \cdot \mathbb{1}_{B^c}(x) \\ &= f(w) \cdot \mathbb{1}_B(x) + f(x) \cdot \mathbb{1}_{B^c}(x). \end{aligned}$$

Since $w \notin B^{\mathsf{c}}$ we get $\delta_w(B^{\mathsf{c}}) = 0$ and thus $\int f(x) \cdot \mathbb{1}_{B^{\mathsf{c}}}(x)\, \delta_w(dx) = 0$. Together we get:

$$\begin{aligned}\int f(x)\, \delta_w(dx) &= \int \left(f(w) \cdot \mathbb{1}_B(x) + f(x) \cdot \mathbb{1}_{B^{\mathsf{c}}}(x)\right)\, \delta_w(dx) \\ &= f(w) \cdot \int \mathbb{1}_B(x)\, \delta_w(dx) + \underbrace{\int f(x) \cdot \mathbb{1}_{B^{\mathsf{c}}}(x)\, \delta_w(dx)}_{=0} \\ &= f(w) \cdot \delta_w(B) \\ &= f(w).\end{aligned}$$

□

2. ***Discrete distributions***. *Consider a discrete probability distribution P supported on the countable subset $\mathcal{W} \subseteq \mathcal{X}$. Let p be its mass function. We then can write the corresponding probability measure P on $\mathcal{X}$ as:*

$$P = \sum_{w \in \mathcal{W}} p(w) \cdot \delta_w.$$

For measurable $A \subseteq \mathcal{X}$ we then have:

$$P(A) = \sum_{w \in \mathcal{W}} p(w) \cdot \delta_w(A) = \sum_{w \in \mathcal{W} \cap A} p(w).$$

Furthermore, we get:

$$\mathbb{E}[f] = \int f(x)\, P(dx) = \int f(x) \sum_{w \in \mathcal{W}} p(w) \cdot \delta_w(dx) = \sum_{w \in \mathcal{W}} f(w) \cdot p(w).$$

A.6. Densities/Derivatives

Definition A.6.1. *Let $(\mathcal{X}, \mathcal{B})$ be a measure space and μ, ν two measures on it. We say that ν has a **density** w.r.t. μ if there exists a non-negative measurable function $f : \mathcal{X} \to [0, \infty]$ such that for all $A \in \mathcal{B}$:*

$$\nu(A) = \int \mathbb{1}_A \cdot f\, d\mu =: \int_A f\, d\mu.$$

*Such a density does not always exist. If a density exists then it is essentially unique, in the sense that two such densities would only differ on a μ-null set. We often use the notation: $f = \frac{d\nu}{d\mu}$ and call it 'the' **density** or **(Radon-Nikodým) derivative** of ν w.r.t. μ.*

Proposition A.6.2. *Let $(\mathcal{X}, \mathcal{B})$ be a measure space and μ, ν, κ three measures on it and $g : \mathcal{X} \to \bar{\mathbb{R}}$ is either a ν-integrable or non-negative measurable function.*

1. *If ν has a density w.r.t. μ then we have:*

$$\int g \, d\nu = \int g \cdot \frac{d\nu}{d\mu} \, d\mu.$$

2. *(Conic) linearity: If κ has a density w.r.t. μ and ν has a density w.r.t. μ and $a, b \geq 0$ then $a \cdot \kappa + b \cdot \nu$ has a density w.r.t. μ and we have:*

$$\frac{d(a \cdot \kappa + b \cdot \nu)}{d\mu}(x) = a \cdot \frac{d\kappa}{d\mu}(x) + b \cdot \frac{d\nu}{d\mu}(x)$$

   for μ-almost-all $x \in \mathcal{X}$.

3. *Chain rule: If ν has a density w.r.t. μ and μ has a density w.r.t. κ then also ν has a density w.r.t. κ and we have:*

$$\frac{d\nu}{d\kappa}(x) = \frac{d\nu}{d\mu}(x) \cdot \frac{d\mu}{d\kappa}(x)$$

   for κ-almost-all $x \in \mathcal{X}$.

4. *Inverse: If ν has a density w.r.t. μ and μ has a density w.r.t. ν then we have:*

$$\frac{d\nu}{d\mu}(x) = \left(\frac{d\mu}{d\nu}(x) \right)^{-1}$$

   for μ-almost-all $x \in \mathcal{X}$. We can make in this context the (somewhat arbitrary) choice to put: $0^{-1} := \infty$.

Definition A.6.3 (Absolute continuity)**.** *Let μ, ν be two measures on a measurable space $(\mathcal{X}, \mathcal{B})$. We say that ν is* ***absolute continuous*** *w.r.t. μ, in symbols:*

$$\nu \ll \mu,$$

if for every $A \in \mathcal{B}$ with $\mu(A) = 0$ also $\nu(A) = 0$ holds, in short, if:

$$\mu(A) = 0 \quad \Longrightarrow \quad \nu(A) = 0.$$

Theorem A.6.4 (Radon-Nikodým, see [Kle20] Cor. 7.34)**.** *Let $(\mathcal{X}, \mathcal{B}, \mu)$ be a σ-finite measure space and ν another measure on $(\mathcal{X}, \mathcal{B})$. Then the following two statements are equivalent:*

1. *ν has a density w.r.t. μ.*

2. *ν is absolute continuous w.r.t. μ.*

Theorem A.6.5 (Besicovitch density theorem, [Fre15] 472D)**.** *Let μ be a Radon measure on $\mathbb{R}^D$ (e.g. any finite or probability measure or the Lebesgue measure, see A.8.1) and $f : \mathbb{R}^D \to \bar{\mathbb{R}}$ be any (locally) μ-integrable function. Then we have for μ-almost-all $x \in \mathbb{R}^D$:*

1. $\lim_{\varepsilon \to 0} \frac{1}{\mu(B_\varepsilon(x))} \int_{B_\varepsilon(x)} f(z)\, \mu(dz) = f(x)$.

2. $\lim_{\varepsilon \to 0} \frac{1}{\mu(B_\varepsilon(x))} \int_{B_\varepsilon(x)} |f(z) - f(x)|\, \mu(dz) = 0$.

Here $B_\varepsilon(x)$ denote the closed balls of radius $\varepsilon > 0$ centered at x (in Euclidean norm). The above, in particular, holds for the density $f = \frac{d\nu}{d\mu}$ of another measure ν w.r.t. μ:

$$\lim_{\varepsilon \to 0} \frac{\nu(B_\varepsilon(x))}{\mu(B_\varepsilon(x))} = \frac{d\nu}{d\mu}(x),$$

for μ-almost-all $x \in \mathbb{R}^D$.

A.7. Conditional Expectation

You may be familiar with the conditional expectation for discrete random variables X, Y:

$$\begin{aligned} \mathbb{E}[X|Y=y] &= \sum_{x \in \mathcal{X}} x \cdot P(X=x|Y=y) \qquad = \sum_{x \in \mathcal{X}} x \cdot \frac{P(X=x, Y=y)}{P(Y=y)} \\ &= \frac{\sum_{x \in \mathcal{X}} x \cdot P(X=x, Y=y)}{P(Y=y)}, \end{aligned}$$

and for real-valued random variables X, Y with positive and continuous joint density $p(x, y)$:

$$\mathbb{E}[X|Y=y] = \int_{\mathcal{X}} x \cdot p(x|Y=y)\, dx = \int_{\mathcal{X}} x \cdot \frac{p(x,y)}{p(y)}\, dx = \frac{\int_{\mathcal{X}} x \cdot p(x,y)\, dx}{p(y)}.$$

The following construction generalizes this notion:

Definition A.7.1 (Conditional expectation)**.** *Let $(\mathcal{W}, P)$ be a probability space and $X : \mathcal{W} \to \mathbb{R}$, $Y : \mathcal{W} \to \mathcal{Y}$ be two random variables with either $\mathbb{E}[|X|] < \infty$ or $X \geq 0$ a.s.*

1. *The **conditional expectation** of X given $Y = y$ is defined via:*

$$\mathbb{E}[X|Y=y] := \mathbb{E}[X_+|Y=y] - \mathbb{E}[X_-|Y=y] \qquad \in \bar{\mathbb{R}},$$

where $X_\pm := \max(\pm X, 0) \geq 0$ and:

$$\mathbb{E}[X_\pm|Y=y] := \frac{dE_\pm}{dP^Y}(y),$$

is the Radon-Nikodym derivative/density w.r.t. P^Y of the following measure on $\mathcal{Y}$:

$$E_\pm(B) := \mathbb{E}\left[X_\pm \cdot \mathbb{1}_B(Y)\right] = \int x \cdot \mathbb{1}_B(y)\, dP^{(X_\pm, Y)}(x, y).$$

One can easily see that $E_\pm \ll P^Y$ and that the densities exist by the Radon-Nikodym theorem.

2. *The **conditional expectation** of X given Y is then the measurable map defined via:*

$$\mathbb{E}[X|Y] : \mathcal{W} \to \bar{\mathbb{R}}, \quad w \mapsto \mathbb{E}[X|Y](w) := \mathbb{E}[X|Y = Y(w)] = \mathbb{E}[X|Y = y]|_{y=Y(w)},$$

i.e. the composition of Y with the measurable map $y \mapsto \mathbb{E}[X|Y = y]$.

Remark A.7.2. *The construction from above also works with a measure μ such that μ^Y is σ-finite (instead of P) since we only need to guarantee the existence of the Radon-Nikodym derivative.*

Notation A.7.3. *Let $\mathcal{W}$, $\mathcal{Z}$, $\mathcal{Y}$ be measurable spaces and $Z : \mathcal{W} \to \mathcal{Z}$ and $Y : \mathcal{W} \to \mathcal{Y}$ be measurable maps. We write:*

$$Z \precsim Y$$

if there exists a measurable function $F : \mathcal{Y} \to \mathcal{Z}$ such that $Z = F \circ Y$; in other words if Z is a deterministic (measurable) function of Y, i.e.: $Z = F(Y)$.
If μ is a measure on $\mathcal{W}$ we also write:

$$Z \precsim_\mu Y$$

if there exists a measurable map F such that $Z = F(Y)$ μ-almost-surely.

Theorem A.7.4. *Let $(\mathcal{W}, P)$ be a probability space and $X, T : \mathcal{W} \to \mathbb{R}$, $Y : \mathcal{W} \to \mathcal{Y}$, $Z : \mathcal{W} \to \mathcal{Z}$ be random variables with $\mathbb{E}[|X|] < \infty$ (or as long as we do not run into the "$\infty - \infty = ?$" problem). Then we have the following properties:*

1. *$\mathbb{E}[X|Y]$ is the unique real valued random variable Z (up to P-null set) such that:*
   a) *$Z \precsim_P Y$ and:*
   b) *for all measurable $B \subseteq \mathcal{Y}$:*

$$\mathbb{E}\left[Z \cdot \mathbb{1}_B(Y)\right] = \mathbb{E}[X \cdot \mathbb{1}_B(Y)].$$

2. *For all real valued random variables $Z \precsim_P Y$ with $\mathbb{E}[|Z \cdot X|] < \infty$ we have:*

$$\mathbb{E}\left[Z \cdot X|Y\right] = Z \cdot \mathbb{E}[X|Y] \quad P\text{-a.s.}$$

3. *Linearity: For all $a, b \in \mathbb{R}$ we have:*

$$\mathbb{E}[a \cdot X + b \cdot T|Y] = a \cdot \mathbb{E}[X|Y] + b \cdot \mathbb{E}[T|Y] \quad P\text{-a.s.}$$

4. *Constants: $\mathbb{E}[1|Y] = 1$ P-a.s.*
5. *Constant maps: If Y is a constant map then: $\mathbb{E}[X|Y] = \mathbb{E}[X]$ P-a.s.*
6. *Independence (see 2.5.23): If $X \perp\!\!\!\perp Y$ then: $\mathbb{E}[X|Y] = \mathbb{E}[X]$ P-a.s.*
7. *Deterministic dependence: If $X \precsim_P Y$ then: $\mathbb{E}[X|Y] = X$ P-a.s.*

8. *Monotonicity: If $X \geq T$ P-a.s. then we have:*

$$\mathbb{E}[X|Y] \geq \mathbb{E}[T|Y] \quad P\text{-a.s.}$$

9. *Jensen inequality: Let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex then we have:*

$$\varphi\left(\mathbb{E}[X|Y]\right) \leq \mathbb{E}[\varphi(X)|Y] \quad P\text{-a.s.}$$

10. *Triangle inequality: $|\mathbb{E}[X|Y]| \leq \mathbb{E}[\,|X|\,|\,Y]$ P-a.s.*

11. *Tower rule: If $Y \precsim Z$ then:*

$$\mathbb{E}\left[\mathbb{E}[X|Y]|Z\right] = \mathbb{E}\left[\mathbb{E}[X|Z]|Y\right] = \mathbb{E}[X|Y] \quad P\text{-a.s.}$$

12. *Tower rule, special case:*

$$\mathbb{E}\left[\mathbb{E}[X|Y]|Y,Z\right] = \mathbb{E}\left[\mathbb{E}[X|Y,Z]|Y\right] = \mathbb{E}[X|Y] \quad P\text{-a.s.}$$

13. *Monotone convergence, dominated convergence, etc. (see literature).*

A.8. The Lebesgue Measure

Definition A.8.1 (The Lebesgue (outer) measure)**.** *The **Lebesgue (outer) measure** λ^D on $\mathbb{R}^D$ is given for subsets $A \subseteq \mathbb{R}^D$ via:*

$$\lambda^D(A) := \inf\left\{ \sum_{n\in\mathbb{N}} \mathrm{vol}^D\left([a^{(n)}, b^{(n)}]\right) \,\middle|\, A \subseteq \bigcup_{n\in\mathbb{N}} [a^{(n)}, b^{(n)}] \right\},$$

where the infimum is running over sequences of D-dimensional cubes:

$$[a^{(n)}, b^{(n)}] = [a_1^{(n)}, b_1^{(n)}] \times \cdots \times [a_D^{(n)}, b_D^{(n)}],$$

with $a^{(n)} = (a_1^{(n)}, \dots, a_D^{(n)})$, $b^{(n)} = (b_1^{(n)}, \dots, b_D^{(n)}) \in \mathbb{R}^D$, $a_d^{(n)} \leq b_d^{(n)}$ for $d = 1, \dots, D$, $n \in \mathbb{N}$, that jointly cover A, where the D-dimensional volume is given by:

$$\mathrm{vol}^D\left([a^{(n)}, b^{(n)}]\right) := (b_1^{(n)} - a_1^{(n)}) \cdots (b_D^{(n)} - a_D^{(n)}), \qquad \mathrm{vol}^D(\emptyset) := 0.$$

Theorem A.8.2 (The Lebesgue measure)**.** *The **Lebesgue measure** λ^D, when restricted to the Borel-σ-algebra of $\mathbb{R}^D$, is the unique measure on $\mathbb{R}^D$ that satisfies:*

$$\lambda^D\left([a,b]\right) = \mathrm{vol}^D\left([a,b]\right),$$

for all D-dimensional cubes $[a,b]$. If the dimension is clear from the context we might just write λ for λ^D.

Theorem A.8.3. *Let λ be the Lebesgue measure on the interval $[a,b]$. Let $f:[a,b]\to\mathbb{R}$ be a Riemann integrable function (e.g. a continuous function) then f is also λ-integrable and we have:*

$$\int_a^b f(x)\,dx = \int_{[a,b]} f(x)\,\lambda(dx).$$

Theorem A.8.4 (Fundamental theorem of calculus)**.** *Let $f:\mathbb{R}\to\mathbb{R}$ be a measurable function such that $\int_{[a,b]} |f|\,d\lambda < \infty$ for $a,b\in\mathbb{R}$. For fixed $c\in\mathbb{R}$ define $F:[c,\infty)\to\mathbb{R}$ via:*

$$F(x) := \int_{(c,x]} f\,d\lambda.$$

Then F is differentiable in λ-almost-all $x\in\mathbb{R}$ and for those points we have:

$$F'(x) = f(x).$$

A.9. Transformation Rules

Theorem A.9.1 (General integral transformation)**.** *Let $(\mathcal{W},\mu)$ be a measure space and $X:\mathcal{W}\to\mathcal{X}$ and $F:\mathcal{X}\to\mathbb{R}$ be measurable. Then we have:*

$$\int F(X)\,d\mu = \int F\,d(X_*\mu),$$

if either side is well-defined. Written in longer form this is:

$$\int F(X(w))\,\mu(dw) = \int F(x)\,(X_*\mu)(dx).$$

Theorem A.9.2 (Push-forward of densities)**.** *Let $(\mathcal{W},\mu)$ be a measure space and ν another measure on $\mathcal{W}$. Let $\varphi:\mathcal{W}\to\mathcal{Y}$ be a measurable mapping such that $\varphi_*\mu$ is σ-finite. If ν has a density w.r.t. μ then the push-forward measure $\varphi_*\nu$ has a density w.r.t. $\varphi_*\mu$ given as follows:*

$$\frac{d(\varphi_*\nu)}{d(\varphi_*\mu)}(y) = \mathbb{E}_\mu\left[\frac{d\nu}{d\mu}\middle|\varphi=y\right] = \text{"}\int \frac{d\nu}{d\mu}(w)\,\mu(dw|\varphi=y)\text{"},$$

for $\varphi_*\mu$-almost-all $y\in\mathcal{Y}$, where the conditional integral $\mathbb{E}_\mu$ is constructed the same way as the conditional expectation but using the σ-finite measure $\varphi_*\mu$.
If, furthermore, φ is a measurable isomorphism then we get:

$$\frac{d(\varphi_*\nu)}{d(\varphi_*\mu)}(y) = \frac{d\nu}{d\mu}(\varphi^{-1}(y))$$

for $\varphi_*\mu$-almost-all $y\in\mathcal{Y}$.

Theorem A.9.3 (Transformation formula for the Lebesgues measure)**.** *Let $\varphi:\mathbb{R}^D\to\mathbb{R}^D$ be a continously differentiable bijection of $\mathbb{R}^D$ (or of open/closed subsets therein)* *with Jacobian $\varphi'(x)$ at point x. Let λ be the Lebesgue measure on $\mathbb{R}^D$. Then $\varphi_*\lambda$ is absolute continuous w.r.t. λ with density given by:*

$$\frac{d(\varphi_*\lambda)}{d\lambda}(y) = |\det \varphi'(\varphi^{-1}(y))|^{-1}$$

for all $y \in \mathbb{R}^D$ (or in that open/closed subset, and $= 0$ outside).

Corollary A.9.4 (Transformation of (probability) densities w.r.t. the Lebesgue measure)**.** *Let the setting be like in A.9.3. Let ν be a (probability) measure on $\mathbb{R}^D$ with (probability) density p w.r.t. λ. Then $\varphi_*\nu$ also has a (probability) density w.r.t. λ, which is then given by:*

$$\frac{d(\varphi_*\nu)}{d\lambda}(y) = \frac{d(\varphi_*\nu)}{d(\varphi_*\lambda)}(y) \cdot \frac{d(\varphi_*\lambda)}{d\lambda}(y) = p(\varphi^{-1}(y)) \cdot |\det \varphi'(\varphi^{-1}(y)|^{-1}.$$

Theorem A.9.5 (A bit more general, [Fre15] Cor. 263F, 262F(b))**.** *Let $\mathcal{X} \subseteq \mathbb{R}^D$ be a measurable set and $\varphi : \mathcal{X} \to \mathbb{R}^D$ an injective Lipschitz function. Let $\mathcal{X}' \subseteq \mathcal{X}$ be the set of points x at which φ has a derivative $\varphi'(x)$ relative to $\mathcal{X}$[89]. Then we have:*

1. *$\mathcal{X} \setminus \mathcal{X}'$ is a λ-null set.*
2. *$|\det \varphi'| : \mathcal{X}' \to [0, \infty)$ is measurable.*
3. *$\varphi(\mathcal{X}) \subseteq \mathbb{R}^D$ is a measurable set.*
4. *$\lambda(\varphi(\mathcal{X})) = \int_{\mathcal{X}} |\det \varphi'(x)|\, d\lambda(x)$.*
5. *For every real-valued function g defined on a subset $\mathcal{Y} \subseteq \varphi(\mathcal{X})$ we have:*

$$\int_{\varphi(\mathcal{X})} g(y)\, d\lambda(y) = \int_{\mathcal{X}} g(\varphi(x)) \cdot |\det \varphi'(x)|\, d\lambda(x),$$

   if either integral is defined in $[-\infty, \infty]$ and provided we interpret $g(\varphi(x)) \cdot |\det \varphi'(x)| := 0$ if $\varphi(x) \notin \mathcal{Y}$ and $|\det \varphi'(x)| = 0$.

Remark A.9.6 (Transformation rule for discrete measures)**.** *Let $\mathcal{X}$ be a measurable space and μ be a discrete (probability) measure on $\mathcal{X}$ supported on the countable discrete subset $\mathcal{W} \subseteq \mathcal{X}$ with mass function given by:*

$$m(x) = \frac{d\mu}{d\#_{\mathcal{W}}}(x),$$

where $\#_{\mathcal{W}}$ is the counting measure w.r.t. $\mathcal{W}$ given by: $\#_{\mathcal{W}}(A) := \#(\mathcal{W} \cap A)$. Let $\varphi : \mathcal{X} \to \mathcal{Y}$ be a measurable map. Then $\varphi_*\mu$ is a discrete measure supported on $\varphi(\mathcal{W})$ with mass function/density:

$$\frac{d\varphi_*\mu}{d\#_{\varphi(\mathcal{W})}}(y) = \sum_{w \in \varphi^{-1}(y) \cap \mathcal{W}} m(w).$$

[89] We say that φ is *differentiable relative to* $\mathcal{X}$ at $x \in \mathcal{X}$ if there exists $\varphi'(x) \in \mathbb{R}^{D \times D}$ such that for every $\epsilon > 0$ there exists a $\delta > 0$ such that for all $y \in \mathcal{X}$ with $\|y - x\| < \delta$ we have that: $\|\varphi(y) - \varphi(x) - \varphi'(x) \cdot (y - x)\| \leq \epsilon \cdot \|y - x\|$. Note that in this definition such a derivative $\varphi'(x)$ does not need to be unique.

Example A.9.7 (Linear transformation of Gaussian distributions).

Example A.9.8 (Density of Chi-square distributions).

A.10. Measure Extension Theorems

Theorem A.10.1 (Measure extension theorem, see [Kle20] Thm. 1.53, Thm. 1.36). *Let $\mathcal{A}$ be a ring (e.g. an algebra) of subsets of a set $\mathcal{X}$. Let $\mu : \mathcal{A} \to [0, \infty)$ be a (finitely) additive set function with $\mu(\emptyset) = 0$ that is also **$\emptyset$-continuous**:*

$$\inf_{n \in \mathbb{N}} \mu(A_n) = 0, \tag{72}$$

for all non-increasing sequences $(A_n)_{n\in\mathbb{N}}$ with $A_n \in \mathcal{A}$ and $A_{n+1} \subseteq A_n$ for all $n \in \mathbb{N}$ and $\bigcap_{n\in\mathbb{N}} A_n = \emptyset$.

Then there exists a unique σ-finite measure $\nu : \sigma(\mathcal{A}) \to [0, \infty]$ such that $\nu(A) = \mu(A)$ for all $A \in \mathcal{A}$.

Theorem A.10.2 (Ionescu-Tulcea extension theorem, see [IT49, Lam87]). *Let I be an arbitrary set (not necessarily countable) and $(\mathcal{X}_i, \mathcal{B}_i)$, $i \in I$, measurable spaces. For subsets $J \subseteq I$ we put:*

$$\mathcal{X}_J := \prod_{j \in J} \mathcal{X}_j, \qquad \mathcal{B}_J := \bigotimes_{j \in J} \mathcal{B}_j, \tag{73}$$

the product space endowed with its product σ-algebra. Now assume that we have a probability measure μ_J on $(\mathcal{X}_J, \mathcal{B}_J)$ for every finite subset $J \subseteq I$ such that:

1. *for every finite subsets $L \subseteq J \subseteq I$ we have: $\mathrm{pr}_{L,*}\mu_J = \mu_L$,*
2. *for every finite subset $J \subseteq I$ and $i \in I \setminus J$ there exists a Markov kernel: $\mu_{i|J} : \mathcal{X}_J \dashrightarrow \mathcal{X}_i$ such that:*

$$\mu_{\{i\} \dot\cup J} = \mu_{i|J} \otimes \mu_J. \tag{74}$$

Then there exists a probability measure μ_I on $(\mathcal{X}_I, \mathcal{B}_I)$ such that for every finite subset $J \subseteq I$ we have:

$$\mathrm{pr}_{J,*}\mu_I = \mu_J. \tag{75}$$

Proof. We first put, with $\mathrm{pr}_J : \mathcal{X}_I \to \mathcal{X}_J$ the canonical projections:

$$\mathcal{A} := \bigcup_{\substack{J \subseteq I \\ \#J < \infty}} \mathrm{pr}_J^* \mathcal{B}_J = \left\{ \mathrm{pr}_J^{-1}(B) \subseteq \mathcal{X}_I \,\middle|\, J \subseteq I, \#J < \infty, B \in \mathcal{B}_J \right\}. \tag{76}$$

Then, per definition, $\mathcal{B}_I = \sigma(\mathcal{A})$. Furthermore, $\mathcal{A}$ is an algebra of subsets of $\mathcal{X}_I$. Indeed, let $A_1, A_2 \in \mathcal{A}$ then $A_l \in \mathrm{pr}_{J_l}^* \mathcal{B}_{J_l}$ for some finite subsets $J_l \subseteq I$, $l = 1, 2$. Then $J := J_1 \cup J_2$

is also a finite subset of I and we have: $A_1, A_2 \in \mathrm{pr}_J^* \mathcal{B}_J$. So, $A_l = \mathrm{pr}_J^{-1}(B_l)$ for some $B_l \in \mathcal{B}_J$, $l = 1, 2$. This then shows that:

$$\begin{aligned} A_l^{\mathsf{c}} &= \mathrm{pr}_J^{-1}(B_l^{\mathsf{c}}) && \in \mathrm{pr}_J^* \mathcal{B}_J \subseteq \mathcal{A}, && (77) \\ A_1 \cup A_2 &= \mathrm{pr}_J^{-1}(B_1 \cup B_2) && \in \mathrm{pr}_J^* \mathcal{B}_J \subseteq \mathcal{A}, && (78) \\ A_1 \cap A_2 &= \mathrm{pr}_J^{-1}(B_1 \cap B_2) && \in \mathrm{pr}_J^* \mathcal{B}_J \subseteq \mathcal{A}. && (79) \end{aligned}$$

It is also clear that: $\mathcal{X}_I, \emptyset \in \mathcal{A}$. So, $\mathcal{A}$ is an algebra of subsets of $\mathcal{X}_I$.

We can now define the set function $\mu : \mathcal{A} \to [0, 1]$ via:

$$\mu(A) := \mu_J(B), \qquad \text{for} \qquad A = \mathrm{pr}_J^{-1}(B), \quad B \in \mathcal{B}_J. \tag{80}$$

This is well-defined because of the condition: $\mathrm{pr}_{L,*}\mu_J = \mu_L$ for finite subsets $L \subseteq J \subseteq I$.

It is also clear that μ is additive. Indeed, if $A_1, A_2 \in \mathcal{A}$ are disjoint then $A_l = \mathrm{pr}_J^{-1}(B_l)$ for some finite subset $J \subseteq I$ and some disjoint $B_l \in \mathcal{B}_J$, $l = 1, 2$. The additivity of μ_J then shows the additivity of μ:

$$\mu(A_1 \,\dot\cup\, A_2) = \mu_J(B_1 \,\dot\cup\, B_2) = \mu_J(B_1) + \mu_J(B_2) = \mu(A_1) + \mu(A_2). \tag{81}$$

To apply the extension theorem A.10.1 it is left to check that μ is $\emptyset$-continuous on $\mathcal{A}$. For this, and, by way of contradiction, consider a non-increasing sequence $A_n \in \mathcal{A}$, $n \in \mathbb{N}$, with $\bigcap_{n\in\mathbb{N}} A_n = \emptyset$ and $\inf_{n\in\mathbb{N}} \mu(A_n) > \epsilon > 0$. We can assume that $A_n = \mathrm{pr}_{J_n}^{-1}(B_n)$ with $B_n \in \mathcal{B}_{J_n}$ with the inclusion of finite subsets: $J_n \subseteq J_{n+1} \subseteq I$ for all $n \in \mathbb{N}$. We totally order the countable set $\bigcup_{n\in\mathbb{N}} J_n$ such that $k < l$ if $k \in J_n$ and $l \in J_{n+1} \setminus J_n$.

We introduce the following abbreviations for $n \in \mathbb{N}$:

$$\mathcal{Y}_n := \mathcal{X}_{J_n \setminus J_{n-1}}, \qquad \mathcal{Y}_{\leq n} := \prod_{l=1}^{n} Y_n, \qquad \mathcal{Y}_{\mathsf{c}} := \mathcal{X}_{I \setminus \bigcup_{n\in\mathbb{N}} J_n}, \tag{82}$$

$$\mu_{n|<n} := \bigotimes_{k \in J_n \setminus J_{n-1}} \mu_{k|\{l \in J_n \,|\, l<k\} \cup J_{n-1}}, \qquad \mu_{\leq n} := \mu_{J_n}. \tag{83}$$

Then $\mathcal{Y} := \mathcal{Y}_{\mathsf{c}} \times \prod_{n\in\mathbb{N}} \mathcal{Y}_n = \mathcal{X}_I$. We also put:

$$h_n(y) := g_n(y_{\leq n}) := \mathbb{1}_{B_n}(y_{\leq n}) = \mathbb{1}_{A_n}(y), \qquad h(y) := \inf_{n\in\mathbb{N}} h_n(y). \tag{84}$$

By assumption, $\bigcap_{n\in\mathbb{N}} A_n = \emptyset$, we have that $h(y) = 0$ for all $y \in \mathcal{Y}$. Since $A_n \subseteq A_{n-1}$ we have for all $n \in \mathbb{N}$:

$$0 = h(y) \leq h_n(y) \leq h_{n-1}(y) \leq 1. \tag{85}$$

We define for $k, n \in \mathbb{N}$:

$$f_n^{(k)}(y_{\leq k}) := \int g_n(y_{k+1:n}, y_{\leq k})\, \mu_{k+1:n|\leq k}(dy_{k+1:n}|y_{\leq k}), \tag{86}$$

$$f^{(k)}(y_{\leq k}) := \inf_{n\in\mathbb{N}} f_n^{(k)}(y_{\leq k}). \tag{87}$$

Note that we then also have for all $k, n \in \mathbb{N}$ and $y \in \mathbb{N}$:

$$0 \leq f^{(k)}(y_{\leq k}) \leq f_n^{(k)}(y_{\leq k}) \leq f_{n-1}^{(k)}(y_{\leq k}) \leq 1. \tag{88}$$

We also put for $k \in \mathbb{N}$:

$$C_{\leq k} := \left\{ y_{\leq k} \in \mathcal{Y}_{\leq k} \,\middle|\, f^{(k)}(y_{\leq k}) > \epsilon \right\}. \tag{89}$$

By the above assumption we have for every $n \in \mathbb{N}$:

$$\begin{aligned} 0 < \epsilon < & \inf_{n\in\mathbb{N}} \mu(A_n) && (90)\\ &= \inf_{n\in\mathbb{N}} \mu_{\leq n}(B_n) && (91)\\ &= \inf_{n\in\mathbb{N}} \mathbb{E}_{\mu_{\leq n}}[g_n] && (92)\\ &= \inf_{n\in\mathbb{N}} \int \int g_n(y_{2:n}, y_1)\, \mu_{2:n|1}(dy_{2:n}|y_1)\, \mu_1(dy_1) && (93)\\ &= \inf_{n\in\mathbb{N}} \int f_n^{(1)}(y_1)\, \mu_1(dy_1) && (94)\\ &= \int \inf_{n\in\mathbb{N}} f_n^{(1)}(y_1)\, \mu_1(dy_1) && (95)\\ &= \int f^{(1)}(y_1)\, \mu_1(dy_1). && (96) \end{aligned}$$

Where the integral and infimum can be interchanged because $f^{(1)}$ is $\mathcal{B}_1$-measurable and the monotone convergence theorem, applied to $1 - f_n^{(1)}$. We see that: $\mu_1(C_1) > 0$. Otherwise, $\int f^{(1)}(y_1)\, \mu_1(dy_1) \leq \epsilon$, which would contradict the above sequence of inequalities.

Now inductively for $k \in \mathbb{N}$ and $y_{\leq k} \in C_{\leq k}$ we have:

$$\begin{aligned} 0 < \epsilon < & f^{(k)}(y_{\leq k}) && (97)\\ &= \inf_{n\in\mathbb{N}} f_n^{(k)}(y_{\leq k}) && (98)\\ &= \inf_{n\in\mathbb{N}} \int f_n^{(k+1)}(y_{k+1}, y_{\leq k})\, \mu_{k+1|\leq k}(dy_{k+1}|y_{\leq k}) && (99)\\ &= \int \inf_{n\in\mathbb{N}} f_n^{(k+1)}(y_{k+1}, y_{\leq k})\, \mu_{k+1|\leq k}(dy_{k+1}|y_{\leq k}) && (100)\\ &= \int f^{(k+1)}(y_{k+1}, y_{\leq k})\, \mu_{k+1|\leq k}(dy_{k+1}|y_{\leq k}). && (101) \end{aligned}$$

This shows that: $\mu_{k+1|\leq k}(C_{\leq k+1}^{y_{\leq k}}|y_{\leq k}) > 0$ for $y_{\leq k} \in C_{\leq k}$. This means that we can inductively construct a $y \in \mathcal{Y}$ with components: $y_1 \in C_1$ and $y_{k+1} \in C_{\leq k+1}^{y_{\leq k}}$ for $k \in \mathbb{N}$, and an arbitrary $y_{\mathsf{c}} \in \mathcal{Y}_{\mathsf{c}}$. This y then satisfies $h_n(y) > \epsilon > 0$ for all $n \in \mathbb{N}$ and thus $h(y) = \inf_{n\in\mathbb{N}} h_n(y) \geq \epsilon > 0$, which lies in contradiction to $h(y) = 0$ for all $y \in \mathcal{Y}$.

This shows that μ is $\emptyset$-continuous. It follows by the extension theorem A.10.1 that μ has a unique extension to a probability measure to $\sigma(\mathcal{A}) = \mathcal{B}_I$. This shows the claim. □

Corollary A.10.3 (Ionescu-Tulcea extension theorem for Markov kernels)**.** *Let I be an arbitrary set (not necessarily countable) and $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$ and $(\mathcal{X}_i, \mathcal{B}_i)$, $i \in I$, measurable spaces. For subsets $J \subseteq I$ we put:*

$$\mathcal{X}_J := \prod_{j \in J} \mathcal{X}_j, \qquad\qquad \mathcal{B}_J := \bigotimes_{j \in J} \mathcal{B}_j, \tag{102}$$

the product space endowed with its product σ-algebra. Now assume that for every finite subset $J \subseteq I$ we are given a Markov kernel:

$$K_J(X_J|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}_J,$$

such that:

1. *for every finite subsets $L \subseteq J \subseteq I$ we have:*

$$K_J(X_L|Z) = K_L(X_L|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}_L,$$

2. *for every finite subset $J \subseteq I$ and $i \in I \setminus J$ there exists a Markov kernel:*

$$K_{i|J}(X_i|X_J, Z) : \mathcal{X}_J \times \mathcal{Z} \dashrightarrow \mathcal{X}_{\{i\} \dot\cup J},$$

   such that:

$$K_{i|J}(X_i|X_J, Z) \otimes K_J(X_J|Z) = K_{\{i\} \dot\cup J}(X_i, X_J|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}_{\{i\} \dot\cup J}.$$

Then there exists a Markov kernel:

$$K(X_I|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}_I,$$

such that for every finite subset $J \subseteq I$ we have:

$$K(X_J|Z) = K_J(X_J|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}_J.$$

Proof. For every $z \in \mathcal{Z}$ we can apply the Ionescu-Tulcea extension theorem A.10.2 separately and get a probability measure $K(X_I|Z = z)$ on $\mathcal{B}_I$ such that for every finite subset $J \subseteq I$ we have:

$$K(X_J|Z = z) = K_J(X_J|Z = z).$$

We are left to check that the map:

$$K(X_I|Z) : \mathcal{Z} \to \mathcal{P}(\mathcal{X}_I), \qquad z \mapsto K(X_I|Z = z),$$

is measurable. By Dynkin's lemma and the definition of the product σ-algebra $\mathcal{B}_I$ this only needs to be checked on sets $A_J \in \mathcal{B}_J$ for finite subsets $J \subseteq I$. Since we have:

$$K(X_I \in \mathrm{pr}_J^{-1}(A)|Z = z) = K(X_J \in A|Z = z) = K_J(X_J \in A|Z = z),$$

and $z \mapsto K_J(X_J \in A|Z = z)$ is measurable for finite $J \subseteq I$, the claim follows. $\square$

Corollary A.10.4 (Kolmogorov extension theorem for Markov kernels). *Let I be an arbitrary set (not necessarily countable) $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$ a measurable space and $(\mathcal{X}_i, \mathcal{B}_i)$, $i \in I$,* ***standard*** *measurable spaces. For subsets $J \subseteq I$ we put:*

$$\mathcal{X}_J := \prod_{j \in J} \mathcal{X}_j, \qquad \mathcal{B}_J := \bigotimes_{j \in J} \mathcal{B}_j, \tag{103}$$

the product space endowed with its product σ-algebra. Now assume that for every finite subset $J \subseteq I$ we are given a Markov kernel:

$$K_J(X_J|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}_J,$$

such that for every finite subsets $L \subseteq J \subseteq I$ we have:

$$K_J(X_L|Z) = K_L(X_L|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}_L.$$

Then there exists a Markov kernel:

$$K(X_I|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}_I,$$

such that for every finite subset $J \subseteq I$ we have:

$$K(X_J|Z) = K_J(X_J|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}_J.$$

Proof. This directly follows from Ionescu-Tulcea extension theorem for Markov kernels A.10.3 and the fact that on standard measurable spaces we always have conditional Markov kernels by the disintegration theorem 2.4.16. ∎

References

[ARSZ05] R. Ayesha Ali, Thomas S. Richardson, Peter Spirtes, and Jiji Zhang. Towards characterizing Markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the 21th Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 10–17, 2005.

[BFPM21] Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *Annals of Statistics*, 49(5):2885–2915, 2021. doi:10.1214/21-AOS2064.

[BJvdV17] Fetsje Bijma, Marianne Jonker, and Aad van der Vaart. *An Introduction to Mathematical Statistics*. Amsterdam University Press, 2017.

[BM18] Stephan Bongers and Joris M. Mooij. From random differential equations to structural causal models: the stochastic case. *arXiv.org preprint*, arXiv:1803.08784v2 [cs.AI], March 2018. URL: https://arxiv.org/abs/1803.08784v2.

[Bol89] Kenneth A. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, Inc., 1989.

[BP94a] A. Balke and J. Pearl. Counterfactual probabilities: computational methods, bounds and applications. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence (UAI-94)*, pages 46–54, 1994.

[BP94b] A. Balke and J. Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the 12th Conference on Artificial Intelligence*, volume 1, pages 230–237. MIT Press, 1994.

[BTP14] E. Bareinboim, J. Tian, and J. Pearl. Recovering from selection bias in causal and statistical inference. In *AAAI*, 2014.

[BvDM21] Tineke Blom, Mirthe M. van Diepen, and Joris M. Mooij. Conditional independences and causal relations implied by sets of equations. *Journal of Machine Learning Research*, 22(178):1–62, 2021. URL: http://jmlr.org/papers/v22/20-863.html.

[BW16] Rabi Bhattacharya and Edward C. Waymire. *A Basic Course in Probability Theory*. Springer, 2nd edition, 2016.

[CB17] J. D. Correa and E. Bareinboim. Causal effect identification by adjustment under confounding and selection biases. In *AAAI*, 2017.

[CD17] Panayiota Constantinou and A. Philip Dawid. Extended conditional independence and applications in causal inference. *The Annals of Statistics*, pages 2618–2653, 2017.

[CE24] Patrick Cheridito and Stephan Eckstein. Optimal transport and Wasserstein distances for causal models. *arXiv.org preprint*, arXiv:2303.14085v2 [math.ST], July 2024. URL: https://arxiv.org/abs/2303.14085v2.

[CH11] Tom Claassen and Tom Heskes. A logical characterization of constraint-based causal discovery. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 135–144, 2011.

[CMH13] Tom Claassen, Joris M. Mooij, and Tom Heskes. Learning sparse causal models is not NP-hard. In Ann Nicholson and Padhraic Smyth, editors, *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, pages 172–181. AUAI Press, 2013. URL: http://auai.org/uai2013/prints/papers/121.pdf.

[Coh13] Donald L. Cohn. *Measure Theory*. Birkhäuser, Boston, USA, 2nd edition, 2013.

[Coo97] Gregory F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1(2):203–224, 1997.

[CTB18] J. D. Correa, J. Tian, and E. Bareinboim. Generalized adjustment under confounding and selection biases. In *AAAI*, 2018.

[Dar53] George Darmois. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut international de statistique*, pages 2–8, 1953.

[Daw79] A. Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979.

[Daw80] A. Philip Dawid. Conditional independence for statistical operations. *The Annals of Statistics*, pages 598–617, 1980.

[Daw01] A. Philip Dawid. Separoids: a mathematical framework for conditional independence and irrelevance. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):335–372, 2001.

[Daw02] A. Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70:161–189, 2002.

[DM83] Claude Dellacherie and Paul-André Meyer. *Probability and Potential B. Theory of Martingales*. Number 72 in North-Holland Mathematics Studies. Elsevier Science, 1983. Translated from French by J. P. Wilson.

[FM17] Patrick Forré and Joris M. Mooij. Markov properties for graphical models with cycles and latent variables. *arXiv.org preprint*, arXiv:1710.08775 [math.ST], 2017. URL: https://arxiv.org/abs/1710.08775.

[FM18] Patrick Forré and Joris M. Mooij. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-2018)*, 2018.

[FM20] Patrick Forré and Joris M. Mooij. Causal calculus in the presence of cycles, latent confounders and selection bias. In *Proceedings of the 35th Annual Conference on Uncertainty in Artificial Intelligence (UAI-2019)*, volume 115, pages 71–80. PMLR, 2020. URL: http://proceedings.mlr.press/v115/forre20a.html.

[For21] Patrick Forré. Transitional conditional independence. *arXiv.org preprint*, arXiv:2104.11547 [math.ST], 2021. URL: https://arxiv.org/abs/2104.11547.

[Fre15] David H. Fremlin. *Measure Theory*, volume 1-6. Torres Fremlin, 2000-2015. URL: https://www1.essex.ac.uk/maths/people/fremlin/mt.htm.

[GP95] D. Galles and J. Pearl. Testing identifiability of causal effects. In *UAI*, 1995.

[Haa43] T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1–12, 1 1943.

[HV06] Y. Huang and M. Valtorta. Pearl's calculus of intervention is complete. In *UAI*, 2006.

[HV08] Y. Huang and M. Valtorta. On the completeness of an identifiability algorithm for semi-Markovian models. *Annals of Mathematics and Artificial Intelligence*, 54(4):363–408, 2008.

[IT49] Cassius T. Ionescu Tulcea. Mesures dans les espaces produits. *Atti Accad. Naz. Lincei Rend*, 7:208–211, 1949.

[Kal17] Olav Kallenberg. *Random Measures, Theory and Applications*, volume 77 of *Probability Theory and Stochastic Modelling*. Springer, 2017.

[Kec95] Alexander S. Kechris. *Classical Descriptive Set Theory*, volume 156 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.

[Kle20] Achim Klenke. *Probability Theory—A Comprehensive Course*. Universitext. Springer, 3rd edition, 2020.

[Lam87] Charles W. Lamb. A Comparison of Methods for Constructing Probability Measures on Infinite Product Spaces. *Canadian Mathematical Bulletin*, 30(3):282–285, 1987.

[Lau24] Steffen Lauritzen. Total variation convergence preserves conditional independence. *arXiv.org preprint*, arXiv:2401.06177 [math.PR], January 2024. URL: https://arxiv.org/abs/2401.06177.

[LDLL90] S.L. Lauritzen, A.P. Dawid, B.N. Larsen, and H.-G. Leimer. Independence properties of directed Markov fields. *Networks*, 20(5):491–505, 1990.

[Man06] Subramani Mani. *A Bayesian Local Causal Discovery Framework*. PhD thesis, University of Pittsburg, March 2006. URL: http://d-scholarship.pitt.edu/10181/.

[MC20] Joris M. Mooij and Tom Claassen. Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI-20)*, volume 124, pages 1159–1168. PMLR, 8 2020. URL: http://proceedings.mlr.press/v124/m-mooij20a/m-mooij20a-supp.pdf.

[Mes12] Franz H. Messerli. Chocolate consumption, cognitive function, and Nobel laureates. *N Engl J. Med.*, 367:1562–1564, 2012. doi:10.1056/NEJMon1211064.

[MMC20] Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020. URL: http://jmlr.org/papers/v21/17-123.html.

[MN98] Charles F. Manski and Daniel S. Nagin. Bounding disagreements about treatment effects. *Sociological Methodology*, 28(1):99–137, 1998.

[MP13] Charles F. Manski and John V. Pepper. Deterrence and the death penalty: Partial identification analysis using repeated cross sections. *Journal of Quantitative Criminology*, 29:123–141, 2013.

[Par05] Kalyanapuram Rangachari Parthasarathy. *Probability Measures on Metric Spaces*, volume 352. American Mathematical Society, 2005.

[PBM16] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B*, 78(5):947–1012, 2016.

[Pea93a] J. Pearl. Aspects of graphical models connected with causality. In *In Proceedings of the 49th Session of the International Statistical Institute*, pages 391–401, 1993.

[Pea93b] J. Pearl. Comment: Graphical models, causality, and intervention. *Statistical Science*, 8(3):266–269, 1993.

[Pea95] J. Pearl. Causal diagrams for empirical research (with discussion). *Biometrika*, 82(4):669–710, 1995.

[Pea09] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

[PJS17] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundation and Learning Algorithms*. MIT Press, 2017.

[PP85] Judea Pearl and Azaria Paz. *Graphoids: A Graph-based Logic for Reasoning about Relevance Relations*. University of California (Los Angeles). Computer Science Department, 1985.

[PP10] J. Pearl and A. Paz. Confounding equivalence in causal inference. In *UAI*, 2010.

[PTKM15] E. Perkovic, J. Textor, M. Kalisch, and M. Maathuis. A complete generalized adjustment criterion. In *UAI*, 2015.

[RERS23] Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov Properties for Acyclic Directed Mixed Graphs. *The Annals of Statistics*, 51(1):334–361, 2023.

[Ric03] Thomas S. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

[RR13a] Thomas S. Richardson and James M. Robins. Single world intervention graphs: A primer. In *Second UAI Workshop on Causal Structure Learning, Bellevue, Washington*, 2013.

[RR13b] Thomas S. Richardson and James M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30), 2013.

[RS02] Thomas S. Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, August 2002.

[SC99] Peter Spirtes and Gregory F. Cooper. An experiment in causal discovery using a pneumonia database. In David Heckerman and Joe Whittaker, editors, *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*, volume R2 of *Proceedings of Machine Learning Research*. PMLR, 1999.

[SdWR10] I. Shpitser, T. J. Van der Weele, and J. M. Robins. On the validity of covariate adjustment for estimating causal effects. In *UAI*, 2010.

[SGS00] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

[SMR95] Peter Spirtes, Christopher Meek, and Thomas S. Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 499–506. Morgan Kaufmann, 1995.

[SMR99] Peter Spirtes, Christopher Meek, and Thomas Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. In Clark Glymour and Gregory F. Cooper, editors, *Computation, Causation, and Discovery*, chapter 6, pages 211–252. MIT Press, Cambridge, MA, 1999.

[SP06a] I. Shpitser and J. Pearl. Identification of conditional interventional distributions. In *UAI*, 2006.

[SP06b] I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *AAAI*, 2006.

[Spi94] P. Spirtes. Conditional independence in directed cyclic graphical models for feedback. Technical Report CMU-PHIL-54, Carnegie Mellon University, 1994.

[Spi95] P. Spirtes. Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 499–506, 1995.

[SPP+05] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308, 2005.

[SW60] Robert H. Strotz and H. O. A. Wold. Recursive vs. nonrecursive systems: An attempt at synthesis. *Econometrica*, 28(2):417–427, 1960.

[Tia02] J. Tian. *Studies in Causal Reasoning and Learning.* PhD thesis, University of California, Los Angeles, 2002.

[Tia04] J. Tian. Identifying conditional causal effects. In *UAI*, 2004.

[TP02] J. Tian and J. Pearl. A general identification condition for causal effects. In *AAAI*, 2002.

[Č82] Nikolai Nikolaevich Čencov. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs.* American Mathematical Society, 1982. translated from Russian.

[vdV98] A. W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, 1998.

[Ver93] Tom S. Verma. Graphical aspects of causal models. Technical Report R-191, Computer Science Department, University of California, Los Angeles, 1993.

[vH48] Joanne Baptista van Helmont. *Ortus medicinae: Id est Initia physicae inaudita. Progressus medicinae novus, in morborum ultionem, ad vitam longam.* Apud Ludovicum Elzevirium, 1648.

[VM19] Philip Versteeg and Joris M. Mooij. Boosting local causal discovery in high-dimensional expression data. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2599–2604, Los Alamitos, CA, USA, 11 2019. IEEE Computer Society. URL: https://doi.ieeecomputersociety.org/10.1109/BIBM47256.2019.8983232, doi:10.1109/BIBM47256.2019.8983232.

[VP90] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In L. Kanal M. Henrion, R. Shachter and J. Lemmer, editors, *Uncertainty in Artificial Intelligence (UAI-90)*, pages 220–227. Association for Uncertainty in AI, 1990.

[Wag20] Stefan Wager. Stats 361: Causal inference. Technical report, Stanford University, 2020. URL: https://web.stanford.edu/~swager/stats361.pdf.

[Wri21] S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

[Wri28] Philip G. Wright. *The Tariff on Animal and Vegetable Oils*, chapter Appendix B, pages 286–319. The MacMillan Company, New York, USA, 1928.

[Zha06] Jiji Zhang. *Causal Inference and Reasoning in Causally Insufficient Systems*. PhD thesis, Carnegie Mellon University, July 2006. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.466.7206&rep=rep1&type=pdf.

[Zha08] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.