Causality

Lecture Notes

Dr. Patrick Forré
Prof. Joris M. Mooij

February 7, 2022

Table of Contents

1. Experimental Causal Discovery

1.1. Types of Correlations

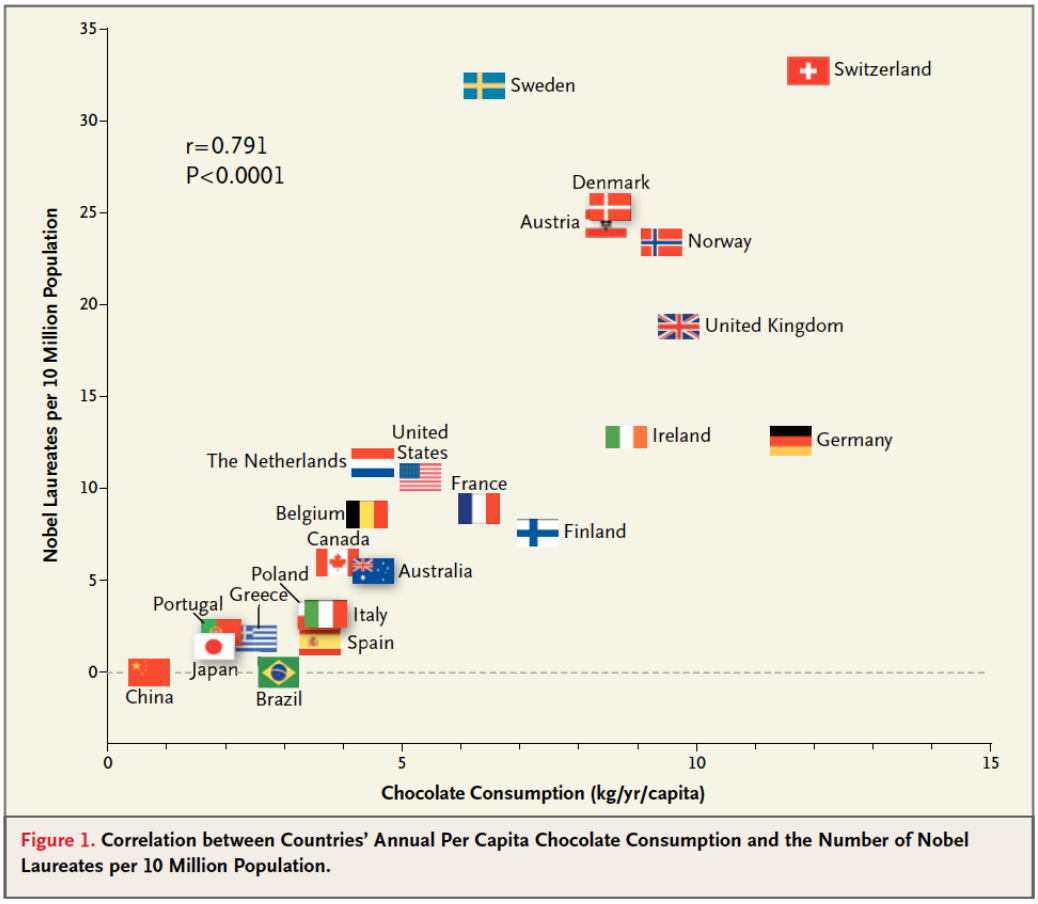


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Figure 1: Correlation between chocolate consumption and Nobel prizes [Mes12].

Explanation 1.1.1. *What conclusions can we make 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 measured 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.

d) Selection bias between N and C: "N and C are actually independent, but the data used was biased, e.g. only Western and Asian countries were considered. Other countries might just be in the upper left or bottom right corners."

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 are confounded: "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.?

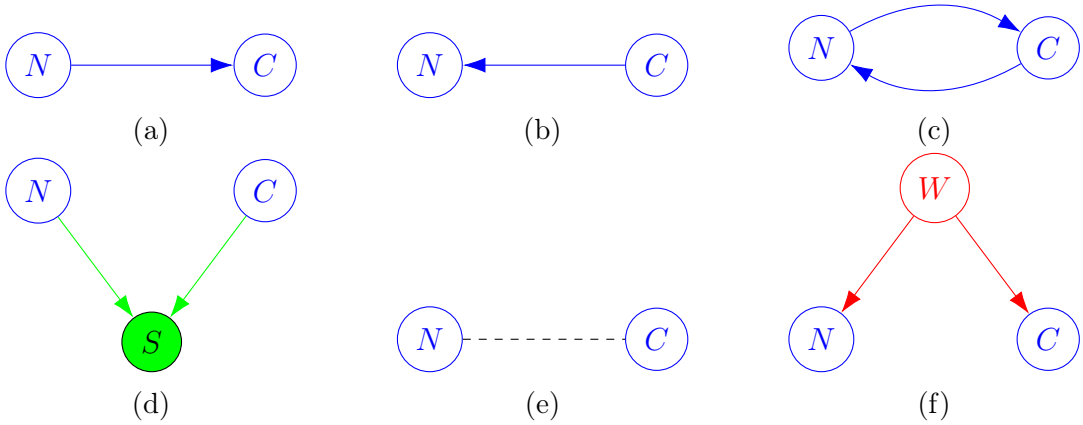


Figure 2: Graphical representations of different correlation inducing scenarios.

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 movement of the the thermometer needle T and the sun S, using 1* 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, using probably many* interventional *data sets, 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—real world definition)**.** *We say that a variable X has* causal effect *onto another variable Y if when* forcing *X to take different values x_0, x_1 then also the distribution of Y changes, in symbols:*

$$\exists x_0, x_1 \in \mathcal{X}: \qquad P(Y|\operatorname{do}(X = x_0)) \neq P(Y|\operatorname{do}(X = x_1)).$$

Remark 1.2.4.

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

$$\exists x_0, x_1 \in \mathcal{X}: \qquad P(Y|X = x_0) \neq P(Y|X = x_1),$$

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

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 can consider 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 variable X has causal influence on outcome variable Y, i.e. we want to estimate the deviation between: $P(Y|\operatorname{do}(X = x_0))$ and $P(Y|\operatorname{do}(X = x_1))$. For this we have test subjects $w_1, \ldots, w_N$. 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 $w_m \in C_0$ from 'control group' the control treatment x_0.*
3. *Measure the outcome y_n for each test subject w_n and estimate the deviation:*

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

4. *Do a statistical test if the deviation D is significantly different from 0.*
5. *If it is significantly different we can conclude a* causal effect *of X onto Y, otherwise not.*

Example 1.3.2. *Example application 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 1.3.3.

1. *An RCT is an 'interventional study' (in contrast to just 'observational study') since we can 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 1.3.4. *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. *if and/how they did doubly/triply blind RCTs (in case they did one),*
3. *otherwise, if and/or how they ruled out other correlation inducing scenarios,*
4. *what bias could have possibly 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. Be creative, create 5 stories!*

Write your findings down 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:*

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

where $F^{-1}(C) := \{w \in \mathcal{W} \mid F(w) \in C\}$ *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 given 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 (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:

$$P(\{x\}) = 0.$$

Now consider 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 e.g. Machine Learning classes.*

3. *Note that, in the discrete case with K classes, one only needs to specify the K values $p(1), \ldots, p(K)$, 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 those values can be derived from $p(w)$ 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 $w \in \mathcal{W}$.*

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 subsets $A \subseteq \mathcal{W}$* 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)**.** *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:*

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

'Measurably isomorphic' means that there is a measurable map 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., PF: cite)**.** *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 2.2.1 (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 2.2.2 (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. finite measure *if $\mu(\mathcal{W}) < \infty$.*

Clearly, every probability measure is finite.

Definition 2.2.3 (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 2.2.4. *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 2.2.5 (See PF: cite)**.** *If $(\mathcal{W}, \mathcal{B}_\mathcal{W})$ is a standard measurable space then also $\mathcal{P}(\mathcal{W})$ and $\mathcal{M}(\mathcal{W})$ are standard measurable spaces (in their 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 1.5. Here we just want to remind the reader of our several different notations, which we will use interchangably during the course:

Notation 2.2.1 (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/Theorem 2.2.1 (The Lebesgue measure)**.** *The* Lebesgue measure λ^D *is the unique measure on $\mathbb{R}^D$ endowed with its Borel-σ-algebra that satsifies:*

$$\lambda^D \left([a_1, b_1] \times \cdots \times [a_D, b_D] \right) = (b_1 - a_1) \cdots (b_D - a_D)$$

for all $a_d, b_d \in \mathbb{R}$, $a_d \leq b_d$, $d = 1, \ldots, D$. If the dimension is clear from the context we might just write λ for λ^D.

2.3. 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 subsets $D \subseteq \mathcal{W}$ a probability value in $[0, 1]$.

Example 2.3.2 (Markov kernels)**.**
1. *any statistical model (i.e. family of model distributions) $\{p_\theta \mid \theta \in \mathcal{F}\}$, 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 (Transition measures and Markov kernels)**.** *Let $\mathcal{T}$, $\mathcal{W}$ be measurable spaces.*

1. *A* (finite[1]) 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 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, a probability measure, resp.

[1] In this course we will only discuss *finite* transition measures and just drop the word "finite" for simplicity in the following.

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

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 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 $t \in \mathcal{T}$[2]).

[2]Note that for a 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 $t \in \mathcal{T}$.

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

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

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 2.2.4. Working out the details is left as an exercise. □

2.3.2. Special Cases of Markov Kernels

Example 2.3.1 (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 $\mathcal{W} = \{w_1, \ldots, w_M\}$ and $\mathcal{T} = \{t_1, \ldots, t_K\}$ 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.

PF: Example, e.g. Normal distribution

Remark 2.3.2 (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 2.3.3 (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 2.3.1 (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 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:*

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

In symbols we abbreviate this as:

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

Remark 2.3.2. *For absolute continuous transition measures $Q(W|T) \ll K(W|T)$ there exists by the Theorem of Radon-Nikodym, see Theorem 1.6.4 or [?] 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 guranteed 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. His approach will be seen to also work for the construction of Doob-Radon-Nikodym derivatives for transition measures. This is why we dedicate his name to the theorem.*

Definition 2.3.3 (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 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 $t \in \mathcal{T}$:

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

but jointly measurably in (w, t).

Lemma 2.3.4. *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 2.3.5 (Null sets)**.** *Let $K(W|T) : \mathcal{T} \to \mathcal{M}(\mathcal{W})$ be a 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 2.3.6 (Essential uniqueness of the Doob-Radon-Nikodym derivative)**.** *Let $\mathcal{T}$, $\mathcal{W}$ be measurable spaces and:*

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

be two transition measures with the two Doob-Radon-Nikodym derivatives g_1 and g_2. 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 is the Doob-Radon-Nikodym derivative essentially unique.

Theorem 2.3.7 (Doob-Radon-Nikodym, see [?] Thm. 58, [?] 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 transition measures. Assume that $\mathcal{W}$ is a standard measurable space[3]. 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 2.3.8. *1. As mentioned in the footnote[3] Theorem 2.3.7 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.*

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

2. *With more technical conditions one could extend Theorem 2.3.7 to work for σ-finite transition measures.*

3. *A special case of Theorem 2.3.7 would be to consider a statistical model $P(X|\Theta)$ that is absolute continuous w.r.t. the Lebesgue measure λ, where $\mathcal{X} = \mathbb{R}^D$. Then one gets a density p w.r.t. λ in arguments: $p(x|\theta)$ that is jointly measurable in (x, θ). Note that λ is not finite, but it is equivalent, in terms of absolute continuity, to a probability measure, e.g. to any Gaussian probability distribution, to which Theorem 2.3.7 then applies.*

Proofs - Theorem of Doob-Radon-Nikodym

Lemma 2.3.9 (Essential uniqueness of the Doob-Radon-Nikodym derivative)**.** *Let $\mathcal{T}$, $\mathcal{W}$ be measurable spaces and:*

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

be two transition measures with the two Doob-Radon-Nikodym derivatives g_1 and g_2. 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 2.3.10 (Existence of the Doob-Radon-Nikodym derivative, see [?] Thm. 58, [?] 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 transition measures. Assume that $\mathcal{W}$ is a standard measurable space[3]*. $Q(W|T) \ll K(W|T)$ 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 $\mathcal{S} = \{[a,b] \,|\, a \leq b, a, b \in \mathbb{Q} \cap [0,1]\}$. We now define the following sequence of finite measurable partitions of $\mathcal{W}$ inductively via:

$$\mathcal{E}_0 := \{\mathcal{W}\}, \qquad \mathcal{E}_n := \{D \setminus D_n, D \cap D_n \,|\, D \in \mathcal{E}_{n-1}\} \setminus \{\emptyset\}, \quad n \in \mathbb{N}.$$

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 $q_D : \mathcal{T} \to \mathbb{R}_{\geq 0}$ 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 $\mathcal{B}_{\mathcal{T}}$-$\mathcal{B}_{\mathbb{R}_{\geq 0}}$-measurable. For $n \in N$ 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 $(\mathcal{B}_{\mathcal{W}} \otimes \mathcal{B}_{\mathcal{T}})$-$\mathcal{B}_{\mathbb{R}_{\geq 0}}$-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 $m \geq n$ 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 [**?**] Lem. 8.10. So we get that G_n^t is a version of $\mathbb{E}_t\left[G_m^t \,|\, \mathcal{B}_n\right]$ 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 [**?**] Ex. 7.39. By the convergence theorem for uniformly integrable martingales, see [**?**] 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 2.3.1 (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 $\mathcal{T} = *$.*

Definition 2.3.2 (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 2.3.3 (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. *$*$ the* constant conditional random variable*:*

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

where $* := \{*\}$ 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 2.4.1 (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 $(w, u) \in \mathcal{W} \times \mathcal{U}$.

Example 2.4.2 (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 2.4.3. *Show that the product of Markov kernels is associative. Under which conditions can we commute Markov kernels in products?*

PF: State Fubini's Theorem!!!!

2.4.3. Composition of Markov Kernels

Definition 2.4.1 (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:

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

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:

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

Remark 2.4.2. *It is clear from the definitions 2.4.1, 2.4.1 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:

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

Remark 2.4.3 (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 2.4.4 (Composition of discrete Markov kernels)**.** *Assume that all the spaces in definition 2.4.1 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 2.4.1 (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 2.4.2. *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 2.4.3. *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 2.4.1, 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 2.4.1 (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 sense that if $Q(X|Y,Z)$ is another such conditional Markov kernel then the set:

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

is a measurable $K(Y|Z)$-null subset of $\mathcal{Y} \times \mathcal{Z}$.

Example 2.4.2 (Conditional Markov kernel for discrete Markov kernels)**.** *Consider a Markov kernel $K(X,Y|Z)$ where all spaces are discrete and k 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:

$$k(x|y,z) := \begin{cases} \frac{k(x,y|z)}{k(y|z)} & \text{if} \quad k(y|z) > 0, \\ k(x|z) & \text{if} \quad k(y|z) = 0.\text{[4]} \end{cases}$$

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 2.4.3 (Conditional probability distributions)**.** *Let X and Y be random variables on domain $(\mathcal{W}, P)$ 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:*[5]

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

Furthermore, these conditional probability distributions are essentially unique.

Proofs - Disintegration of Markov Kernels

Remark 2.4.4 (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 2.3.7, 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

[4] 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.

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

Doob-Radon-Nikodym derivative $K(X \leq x|Y = y, Z = z)$ *for rational points* $x \in \mathbb{Q}$ *and then for general* $x \in \mathbb{R}$ *put:*

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

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:*

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

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

Theorem 2.4.5 (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.* $\mathcal{Y}$ *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:

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

Since also $\mathcal{B}_\mathcal{Y}$ is countably generated, by Doob-Radon-Nikodym, see Theorem 2.3.7, 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:

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

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 [**?**] Cor. 6.26) we get:

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

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 $\mathcal{X}_{\mathbb{Q}} \times D^{\mathsf{c}}$, 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:

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

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 $(x,y,z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$:

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

It is clear that F is again jointly measurable in (y,z) for fixed x and agrees with G on $\mathcal{X}_{\mathbb{Q}} \times D^{\mathsf{c}}$ 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 probability measures in X the system $\mathcal{D}$ is closed under countable disjoint unions and 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 [**?**] 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 2.4.6 (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, $\mathcal{X}$ 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 derivatives of the same transition measure $M_A(Y|Z)$ given by:

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

The uniqueness statement then follows from that of Doob-Radon-Nikodym derivatives, see Lemma 2.3.9. 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. $\mathcal{B}_{[0,1]} = \sigma(\{[0,c] \,|\, c \in [0,1] \cap \mathbb{Q}\})$. 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:

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

One easily sees that $\mathcal{D}$ is closed under complements, countable disjoint unions and contains $\mathcal{X} \in \mathcal{D}$. 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 equality and thus a measurable $K(Y|Z)$-null set. □

2.5. Conditional Independence

2.5.1. Independence of Random Variables

2.5.2. Conditional Independence of Random Variables

2.5.3. Conditional Independence of Conditional Random Variables

In this section we will introduce the notion of conditional independence for conditional random variables from [For21]. It generalizes and unifies stochastic conditional independence and some forms of functional conditional independence. Other versions of extended conditional independence can be found in [Daw79,Daw80,Daw01,CD17,RERS17, FM20].

Definition 2.5.1 (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, Y, Z with common domain $\mathcal{W} \times \mathcal{T}$ and codomains $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, resp. 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 $K(X, Y, Z|T)$.
For the equation above to hold it is sufficient to check that for all $t \in \mathcal{T}$ and all measurable $A \subseteq \mathcal{X}$, $B \subseteq \mathcal{Y}$, $C \subseteq \mathcal{Z}$ one has that:

$$\mathbb{E}_t \left[\mathbb{1}_A(X_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(Z_t) \right] = \mathbb{E}_t \left[Q(X \in A|Z_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(Z_t) \right],$$

where the expectation value $\mathbb{E}_t$ is w.r.t. $K(W|T = t)$, which is short notation for:

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

Special case:

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

Remark 2.5.2 (Essential uniqueness)**.** *The Markov kernel $Q(X|Z)$ appearing in the conditional independence $X \perp\!\!\!\perp_{K(W|T)} Y \,|\, Z$ in definition 2.5.1 is then a version of a conditional Markov kernel $K(X|Y, Z, T)$ and is thus essentially unique in the sense of 2.4.6.*

Notation 2.5.3. *The Markov kernel $Q(X|Z)$ appearing in the conditional independence $X \perp\!\!\!\perp_{K(W|T)} Y \mid Z$ is essentially unique by 2.5.2 and we can write it as:*

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

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

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

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

Remark 2.5.4 (Conditional independence includes conditional independence from T)**.** *We have the equivalence:*

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

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

Proof.

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

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 $\delta(T \in \mathcal{T}|T) = 1$. □

Example 2.5.5 (Conditional independence for discrete conditional random variables)**.** *Let the situation be like in definition 2.5.1 and assume all spaces to be (finite) 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:

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

This is the case if, for example, $k(y,z|t) > 0$ for all values and the quotient function:

$$\tilde{q}(x|y,z,t) := \frac{k(x,y,z|t)}{k(y,z|t)}$$

is - as a function - independent of the arguments y and t.

Remark 2.5.6 (Conditional independence for random variables)**.** *We recover the conditional independence for random variables by taking $\mathcal{T}$ to be the one-point space $* = \{*\}$. Then $(\mathcal{W}, P)$ is a probability space and X, Y, Z random variables. We then get:*

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

If such an $Q(X|Z)$ exists it is clearly a regular conditional probability distribution of $P(X,Z)$ conditioned on Z. In suggestive notations:

$$P(X|\not{Y}, Z) = P(X|Z).$$

By theorem 2.4.1 such a conditional probability distributions always exists and is essentially unique in case $\mathcal{X}$ and $\mathcal{Z}$ are standard measurable spaces. So we could write:

$$P(X|\not{Y}, Z) = P(X|Z).$$

So for the case of standard measurable spaces $\mathcal{X}$, $\mathcal{Z}$ we get:

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

If all three spaces $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are standard one can show the equivalences:

$$\begin{aligned} X \underset{P}{\perp\!\!\!\perp} Y \mid Z \quad &\Longleftrightarrow \quad P(X,Y|Z) = P(X|Z) \otimes P(Y|Z) \quad P(Z)\text{-a.s.} \\ &\Longleftrightarrow \quad P(X|Y,Z) = P(X|Z) \quad P(Y,Z)\text{-a.s.} \\ &\Longleftrightarrow \quad \forall A \in \mathcal{B}_{\mathcal{X}} : \mathbb{E}[\mathbb{1}_A(X)|Y,Z] = \mathbb{E}[\mathbb{1}_A(X)|Z] \quad P\text{-a.s.} \end{aligned}$$

Lemma 2.5.7 (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:*

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

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.
Recall that we write $F \precsim H$ if there exists a measurable mapping $\varphi : \mathcal{H} \to \mathcal{F}$ such that $F = \varphi \circ H$.
Then we have the equivalence[6]*:*

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

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.

[6]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 $\varphi : H(\mathcal{W} \times \mathcal{T}) \to \mathcal{F}$ such that $F = \varphi \circ H$, but this moves problems elsewhere.

Proof. " $\Longrightarrow$ ": 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:

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

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 2.5.8 (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 2.5.9 (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:

$$K(X|\not{T}, Y, T) = K(X|Y, T).$$

Example 2.5.10 (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/or) labels Y are fully captured already by S.

Theorem 2.5.11. *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. $X \underset{K(W|T)}{\perp\!\!\!\perp} T, Y \mid Z$,

3. $X \underset{K(W|T)}{\perp\!\!\!\perp} T \mid Z$ *and* $K(X, Y, Z|T) = P(X|Z) \otimes K(Y, Z|T)$.

4. $X \underset{K(W|T)}{\perp\!\!\!\perp} T \mid Z$ *and for every* $t \in \mathcal{T}$ *we have:* $X_t \underset{K(W|T=t)}{\perp\!\!\!\perp} Y_t \mid Z_t$.

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

Proof. 3. $\implies$ 1. is clear by definition.
1. $\iff$ 2.: by 2.5.4.
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 the factorization:

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

This means that for every $t \in \mathcal{T}$ and every measurable $A \subseteq \mathcal{X}$, $C \subseteq \mathcal{Z}$ we have:

$$\mathbb{E}_t \left[\mathbb{1}_A(X_t) \cdot \mathbb{1}_C(Z_t)\right] = \mathbb{E}_t \left[P(X \in A|Z_t) \cdot \mathbb{1}_C(Z_t)\right],$$

where the expectation $\mathbb{E}_t$ is w.r.t. $K(W|T = t)$. This shows that $P(X \in A|Z_t)$ is a version of $\mathbb{E}_t[\mathbb{1}_A(X_t)|Z_t]$ for every $t \in \mathcal{T}$, by the defining properties of conditional expectation.

By the assumption $X_t \perp\!\!\!\perp_{K(W|T=t)} Y_t \,|\, Z_t$ we then have for every fixed $t \in \mathcal{T}$ and measurable $A \subseteq \mathcal{X}$:

$$\mathbb{E}_t[\mathbb{1}_A(X_t)|Y_t, Z_t] = \mathbb{E}_t[\mathbb{1}_A(X_t)|Z_t] = P(X \in A|Z_t) \qquad K(W|T=t)\text{-a.s.}$$

By the defining properties of conditional expectation for $\mathbb{E}_t[\mathbb{1}_A(X_t)|Y_t, Z_t]$ we then get that for every measurable $A \subseteq \mathcal{X}$, $B \subseteq \mathcal{Y}$, $C \subseteq \mathcal{Z}$:

$$\mathbb{E}_t\left[\mathbb{1}_A(X_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(Z_t)\right] = \mathbb{E}_t\left[P(X \in A|Z_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(Z_t)\right].$$

Since this holds for every $t \in \mathcal{T}$ we get:

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

□

Corollary 2.5.12. *If $\mathcal{X}$, $\mathcal{Z}$ are standard measurable spaces then we have the equivalence:*

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

Proof. This directly follows from theorem 2.5.11 4. with (Z, T) in the role of Z and remark 2.5.9 to get the first part of 4. In suggestive symbols:

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

□

2.6. Separoid Axioms for Conditional Independence

Definition/Theorem 2.6.1 ((Asymmetric) separoid axioms for conditional independence[7]). *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.
Recall that 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:

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

[7] The symmetric separoid axioms are due to A.P. Dawid [Daw01] and the similar graphoid axioms due to J. Pearl and A. Paz [PP85].

b) T-Restricted Right Redundancy[8]*:*

$X \perp\!\!\!\perp *|Z, T$ *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:

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

f) Left Weak Union[8]*:*

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

g) Right Weak Union:

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

h) Left Contraction:

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

i) Right Contraction:

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

j) Right Cross Contraction:

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

k) Flipped Left Cross Contraction:

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

In particular, we have the equivalences:

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

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

We also get:

l) T-Restricted Symmetry[8]*:*

[8](Only) 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.

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

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:

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

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:

$$\mathbb{E}_t \left[Q(X \in A|Z_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(Z_t) \right] = \mathbb{E}_t \left[\mathbb{1}_A(X_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(Z_t) \right]$$

where the expectation value $\mathbb{E}_t$ is w.r.t. $K(W|T = t)$.
We abbreviate $\perp\!\!\!\perp := \perp\!\!\!\perp_{K(W|T)}$ in the following.

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

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

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

$$\mathbb{E}_t \left[\delta_\varphi(U \in D|X_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(X_t) \right] = \mathbb{E}_t \left[\mathbb{1}_D(U_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(X_t) \right].$$

In suggestive symbols:

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

□

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 2.4.1). This already shows the claim. In suggestive symbols:

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

□

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

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

Proof. We use the same $Q(X|Z)$. Then we get by assumption indicated by (!):

$$\begin{aligned}
&\mathbb{E}_t\left[Q(X \in A|Z_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_D(T_t) \cdot \mathbb{1}_C(Z_t)\right] \\
&= \mathbb{E}_t\left[Q(X \in A|Z_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_D(t) \cdot \mathbb{1}_C(Z_t)\right] \\
&= \mathbb{E}_t\left[Q(X \in A|Z_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(Z_t)\right] \cdot \mathbb{1}_D(t) \\
&\overset{(!)}{=} \mathbb{E}_t\left[\mathbb{1}_A(X_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(Z_t)\right] \cdot \mathbb{1}_D(t) \\
&= \mathbb{E}_t\left[\mathbb{1}_A(X_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(Z_t) \cdot \mathbb{1}_D(t)\right] \\
&= \mathbb{E}_t\left[\mathbb{1}_A(X_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_D(T_t) \cdot \mathbb{1}_C(Z_t)\right].
\end{aligned}$$

In suggestive symbols:

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

□

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. Then we use the marginal Markov kernel $Q(U|Z)$. We then get by assumption (!):

$$\begin{aligned}
&\mathbb{E}_t\left[Q(U \in D|Z_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(Z_t)\right] \\
&= \mathbb{E}_t\left[Q(X \in \mathcal{X}, U \in D|Z_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(Z_t)\right] \\
&\overset{(!)}{=} \mathbb{E}_t\left[\mathbb{1}_{\mathcal{X}}(X_t) \cdot \mathbb{1}_D(U_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(Z_t)\right] \\
&= \mathbb{E}_t\left[\mathbb{1}_D(U_t) \cdot \mathbb{1}_B(Y_t) \cdot \mathbb{1}_C(Z_t)\right].
\end{aligned}$$

In suggestive symbols:

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

□

Lemma 2.6.6 (Right Decomposition)**.**

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

Proof. Let $Q(X|Z)$ be given from the left. We then have by assumption (!):

$$\begin{aligned}
&\mathbb{E}_t\left[Q(X \in A|Z_t) \cdot \mathbb{1}_D(U_t) \cdot \mathbb{1}_C(Z_t)\right] \\
&= \mathbb{E}_t\left[Q(X \in A|Z_t) \cdot \mathbb{1}_{\mathcal{Y}}(Y_t) \cdot \mathbb{1}_D(U_t) \cdot \mathbb{1}_C(Z_t)\right] \\
&\overset{(!)}{=} \mathbb{E}_t\left[\mathbb{1}_A(X_t) \cdot \mathbb{1}_{\mathcal{Y}}(Y_t) \cdot \mathbb{1}_D(U_t) \cdot \mathbb{1}_C(Z_t)\right] \\
&= \mathbb{E}_t\left[\mathbb{1}_A(X_t) \cdot \mathbb{1}_D(U_t) \cdot \mathbb{1}_C(Z_t)\right].
\end{aligned}$$

In suggestive symbols:

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

□

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

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

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 2.4.1).
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 $K(X, U|\cancel{T,Y}, Z)$.

□

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

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

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)**.**

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

Proof. By assumption we have the two factorizations:

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

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)**.**

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

Proof. By assumption we have the two factorizations:

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

with some Markov kernels $Q(X|U, Z)$, $P(X|Z)$.
Marginalizing out Y we get the equalities:

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

By the essential uniqueness (see lemma 2.4.6) 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:

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

□

Lemma 2.6.11 (Right Cross Contraction)**.**

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

Proof. By assumption we have the two factorizations:

$$K(X, Y, U, Z|T) = Q(X|U, Z) \otimes K(Y, U, Z|T), \tag{1}$$
$$K(X, U, Z|T) = P(U|Z) \otimes K(X, Z|T), \tag{2}$$

with some Markov kernels $Q(X|U, Z)$, $P(U|Z)$.
We then define the Markov kernel:

$$R(X, U|Z) := Q(X|U, Z) \otimes P(U|Z). \tag{3}$$

We will now show that its marginal:

$$R(X|Z) = Q(X|U, Z) \circ P(U|Z). \tag{4}$$

will satisfy the claim.
If we marginalize out Y from equation 1 we get:

$$K(X, U, Z|T) = Q(X|U, Z) \otimes K(U, Z|T). \tag{5}$$

Equating equations 2 and 5 gives:

$$P(U|Z) \otimes K(X, Z|T) = K(X, U, Z|T) = Q(X|U, Z) \otimes K(U, Z|T). \tag{6}$$

Marginalizing out X in equation 6 on both sides gives:

$$K(U, Z|T) = P(U|Z) \otimes K(Z|T). \tag{7}$$

If we now plug equation 7 into 5 then we get:

$$K(X, U, Z|T) = Q(X|U, Z) \otimes P(U|Z) \otimes K(Z|T) \tag{8}$$
$$\overset{3}{=} R(X, U|Z) \otimes K(Z|T). \tag{9}$$

If we marginalize out U in equation 9 and use definition 4 we arrive at:

$$K(X, Z|T) = R(X|Z) \otimes K(Z|T). \tag{10}$$

We now get:

$$\begin{aligned} Q(X|U, Z) \otimes K(U, Z|T) &\overset{5}{=} K(X, U, Z|T) && (11)\\ &\overset{2}{=} P(U|Z) \otimes K(X, Z|T) && (12)\\ &\overset{10}{=} P(U|Z) \otimes R(X|Z) \otimes K(Z|T) && (13)\\ &= R(X|Z) \otimes P(U|Z) \otimes K(Z|T) && (14)\\ &\overset{7}{=} R(X|Z) \otimes K(U, Z|T). && (15) \end{aligned}$$

By the essential uniqueness (see lemma 2.4.6) of such a factorization we get that for every $A \in \mathcal{B}_{\mathcal{X}}$:

$$Q(X \in A|U, Z) = R(X \in A|Z) \qquad K(U, Z|T)\text{-a.s.} \tag{16}$$

The same equation then holds also $K(Y, U, Z|T)$-a.s. (by ignoring the non-occurring argument y). Plugging 16 back into the equation 1 we get:

$$\begin{aligned} K(X, Y, U, Z|T) &= Q(X|U, Z) \otimes K(Y, U, Z|T), && (17)\\ &= R(X|Z) \otimes K(Y, U, Z|T). && (18) \end{aligned}$$

This shows the claim.
In suggestive symbols:

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

□

Lemma 2.6.12 (Flipped Left Cross Contraction)**.**

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

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:

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

□

Lemma 2.6.13 (T-Restricted Symmetry)**.** *Let $\mathcal{Y}$ and $\mathcal{Z}$ be standard measurable spaces. Then:*

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

Proof. This follows from Flipped Left Cross Contraction with $U = *$ and (Z,T) for Z:

$$(X \perp\!\!\!\perp Y \,|\, Z,T) \quad \wedge \quad (Y \perp\!\!\!\perp * \,|\, Z,T) \implies Y \perp\!\!\!\perp X \,|\, Z,T,$$

together with T-Restricted Right Redundancy:

$$Y \perp\!\!\!\perp * \,|\, Z,T.$$

In suggestive symbols:

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

□

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:*

$$\begin{array}{rcl} \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) & \to & \mathcal{P}(\mathcal{X} \times \mathcal{Y}), \\ (P,Q) & \mapsto & P \otimes Q, \end{array}$$

   is measurable.

4. If $g : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ is measurable then the map:

$$\begin{array}{rcl} \mathcal{P}(\mathcal{X}) \times \mathcal{Y} & \to & \mathcal{P}(\mathcal{Z}), \\ (P, y) & \mapsto & g_*(P \otimes \delta_y) \\ & = & (C \mapsto P(\{x \in \mathcal{X} \,|\, g(x, y) \in C\})) \end{array}$$

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 $\mathcal{X}$ 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, so w.l.o.g. $\mathcal{X} = \bar{\mathbb{R}}$ endowed with the Borel σ-algebra). Let $K(X|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}$ be a Markov kernel, $P(E)$ be the uniform distribution on $\mathcal{E} := [0, 1]$ (i.e. λ) and:*

$$R(e|z) := \inf \{\tilde{x} \in \mathcal{X} \,|K(X \leq \tilde{x}|z) \geq e\},$$

the (conditional) quantile function (aka inverse cumulative distribution function) *of $K(X|Z)$. Then we can write $K(X|Z)$ as the push-forward of the constant uniform one:*

$$K(X|Z) = \delta(R|E, Z) \circ P(E).$$

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 random variable X can be represented as 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 2.7.4. *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, so w.l.o.g. $\mathcal{X} = \bar{\mathbb{R}}$ endowed with the Borel σ-algebra). Let $K(X|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}$ be a Markov kernel. Furthermore, let $\mathcal{U} := [0, 1]$ and $K(U)$ be the uniform distribution/Markov kernel on $\mathcal{U}$. We write:*

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

We then define the interpolated (conditional) cumulative distribution function and its corresponding quantile function:

$$\begin{aligned} F(x; u|z) &:= K(X < x|Z = z) + u \cdot K(X = x|Z = z) \\ R(e|z) &:= \inf \{\tilde{x} \in \mathcal{X} \,|F(\tilde{x}; 1|z) \geq e\}. \end{aligned}$$

We then consider the conditional random variables X, U, Z, E given via:

$$\begin{array}{rccl} X: & \mathcal{X}\times\mathcal{U}\times\mathcal{Z} & \to & \mathcal{X}, \\ & (x,u,z) & \mapsto & x, \\ U: & \mathcal{X}\times\mathcal{U}\times\mathcal{Z} & \to & \mathcal{U}, \\ & (x,u,z) & \mapsto & u, \\ Z: & \mathcal{X}\times\mathcal{U}\times\mathcal{Z} & \to & \mathcal{Z}, \\ & (x,u,z) & \mapsto & z, \\ E: & \mathcal{X}\times\mathcal{U}\times\mathcal{Z} & \to & \mathcal{E}:=[0,1], \\ & (x,u,z) & \mapsto & F(x;u|z). \end{array}$$

Then for all $e \in \mathcal{E}$ and $z \in \mathcal{Z}$ we have:

$$K(E \leq e | Z = z) = e,$$

which is independent of z, which means that:

$$E \underset{K(U,X|Z)}{\perp\!\!\!\perp} Z,$$

with uniform distribution $K(E) = K(E|Z)$ on $[0,1]$. Furthermore, we have:

$$X = R(E|Z) \quad K(U,X|Z)\text{-a.s.}.$$

So that means that X in the Markov kernel $K(X|Z)$ can be replaced by a deterministic map of Z and independent uniformly distributed noise E.
In particular, we get that with $K(E)$ the uniform distribution on $[0,1]$ from above that:

$$K(X|Z) = \delta(R|E,Z) \circ K(E).$$

Proofs - Deterministic Representation of Markov Kernels In this note we generalize a few folklore results via now standard techniques that were introduced in [Dar53].

Lemma 2.7.5. *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:*

$$t \leq F(x) \iff R(t) \leq x.$$

In particular, we have $F(R(t)) \geq t$, thus $R(t) \in F^{-1}([t,1])$ the minimal element. We also have $R(F(x)) \leq x$, 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: $q(x_1) < F_+(x_1) \leq F_-(x_2) < q(x_2)$. 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:

$$F(R(t)) = \inf_{n\in\mathbb{N}} F(x_n) \geq t.$$

It follows that $F(R(t)) \geq t$ 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 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 $\sup_{n\in\mathbb{N}} R(t_n) \leq R(t)$. On the other hand we have:

$$t = \sup_{n\in\mathbb{N}} t_n \leq \sup_{n\in\mathbb{N}} F(R(t_n)) \leq F(\sup_{n\in\mathbb{N}} R(t_n)),$$

implying: $\sup_{n\in\mathbb{N}} R(t_n) \in F^{-1}([t,1])$ and thus $\sup_{n\in\mathbb{N}} R(t_n) \geq R(t)$, leading to equality, which shows the claim.
For any $x \in \bar{\mathbb{R}}$ we have the implication:

$$x \geq R(t) \implies F(x) \geq F(R(t)) \geq t.$$

For any $x \in \bar{\mathbb{R}}$ and any $t \in [0,1]$ we have the implications:

$$\begin{aligned} t \leq F(x) &\iff F(x) \in [t,1] \\ &\iff x \in F^{-1}([t,1]) \\ &\implies x \geq \inf F^{-1}([t,1]) = R(t). \end{aligned}$$

Together this shows for any $x \in \bar{\mathbb{R}}$ and $t \in [0,1]$ the equivalence:

$$t \leq F(x) \iff R(t) \leq x.$$

Since $F(x) \leq F(x)$ we get $R(F(x)) \leq x$ 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 $x \geq R(F(x))$ and $F(R(t)) \geq t$ to conclude:

$$x \geq R(F(x)) = R(F(R(t))) \geq R(t) = x,$$

showing equality, and that:

$$R \circ F \circ R = R.$$

Similarly for $t = F(x)$ we get:

$$t \leq F(R(t)) = F(R(F(x))) \leq F(x) = t,$$

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:

$$\begin{aligned}(R_*\lambda)([-\infty, x]) &= \lambda(R^{-1}([-\infty, x])) \\ &= \lambda(t \in [0,1] \,|\, R(t) \leq x) \\ &= \lambda(t \in [0,1] \,|\, t \leq F(x)) \\ &= \lambda([0, F(x)]) \\ &= F(x) \\ &= P([-\infty, x]).\end{aligned}$$

It follows that: $R_*\lambda = P$. □

Lemma 2.7.6. *Let the notation be like in 2.7.5. 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 $x < \tilde{x}$ then $[-\infty, x] \subseteq [-\infty, \tilde{x})$ and thus $F_1(x) \leq F_0(\tilde{x})$.
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 2.7.5 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 2.7.7. *Let the notations be like in 2.7.5 and 2.7.6. 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:

$$\begin{array}{rccl} E & : & \bar{\mathbb{R}} \times [0,1] & \to & [0,1], \\ & & (x,u) & \mapsto & P([-\infty, x)) + u \cdot P(\{x\}), \end{array}$$

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: $X(x,u) := x$, and 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 2.7.6 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:

$$\begin{aligned}
\bar{P}(E < \tilde{e}) &= \bar{P}(E < F_0(x)) \\
&= \bar{P}(E < F_0(x)) \\
&= \bar{P}(E < \sup_{\tilde{x}<x} F_1(\tilde{x})) \\
&= \sup_{\tilde{x}<x} \bar{P}(E \leq F_1(\tilde{x})) \\
&\overset{(*)}{=} \sup_{\tilde{x}<x} F_1(\tilde{x}) \\
&= F_0(x) \\
&= \tilde{e}.
\end{aligned}$$

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 $p(x) = 0$, the first row 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. □

Theorem 2.7.8. *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, so w.l.o.g. $\mathcal{X} = \bar{\mathbb{R}}$ endowed with the Borel σ-algebra). Let $K(X|Z) : \mathcal{Z} \dashrightarrow \mathcal{X}$ be a Markov kernel. Furthermore, let $\mathcal{U} := [0, 1]$ and $K(U)$ be the uniform distribution/Markov kernel on $\mathcal{U}$. We write:*

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

Also put:

$$\begin{aligned}
F(x; u|z) &:= K(X < x|Z = z) + u \cdot K(X = x|Z = z) \\
R(e|z) &:= \inf\left\{\tilde{x} \in \mathcal{X} \,|F(\tilde{x}; 1|z) \geq e\right\}.
\end{aligned}$$

Let $E := F(X; U|Z)$. We consider X, U, Z, E as the measurable maps:

$$\begin{array}{rccl}
X: & \mathcal{X} \times \mathcal{U} \times \mathcal{Z} & \to & \mathcal{X}, \\
& (x, u, z) & \mapsto & x, \\
U: & \mathcal{X} \times \mathcal{U} \times \mathcal{Z} & \to & \mathcal{U}, \\
& (x, u, z) & \mapsto & u, \\
Z: & \mathcal{X} \times \mathcal{U} \times \mathcal{Z} & \to & \mathcal{Z}, \\
& (x, u, z) & \mapsto & z, \\
E: & \mathcal{X} \times \mathcal{U} \times \mathcal{Z} & \to & \mathcal{E} := [0, 1], \\
& (x, u, z) & \mapsto & F(x; u|z).
\end{array}$$

Then for all $e \in \mathcal{E}$ and $z \in \mathcal{Z}$ we have:

$$K(E \leq e | Z = z) = e.$$

Furthermore, we have:

$$X = R(E|Z) \quad K(U, X|Z)\text{-a.s.}.$$

These two equations imply that for every distribution $P(Z)$, putting $P(U, X, Z) := K(U) \otimes K(X|Z) \otimes P(Z)$, we have:

$$X = R(E|Z) \quad P\text{-a.s.},$$

and that E is uniformly distributed on $\mathcal{E} = [0, 1]$ and P-independent of Z.

Proof. After the measurabilities are checked the statement directly follows from 2.7.7 by applying it for every z separately. □

Corollary 2.7.9. *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. Furthermore, E can be constructed via a deterministic measurable function in X and Z and (uniformly distributed) independent noise U (on $[0, 1]$).*

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 $g(e, z) := R(e|z)$ and E. □

2.8. Markov Kernels from Structural Causal Models

PF: Move this section to SCM part!!!

In this section we will see how Markov kernels appear naturally in the theory of SCMs.

Definition 2.8.1. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ be an SCM with solution $(X_{V \cup W}^{\mathrm{do}(x_J)})_{x_J \in \mathcal{X}_J}$ (see Definition 5.6.5). Then*

$$K : \mathcal{B}_{\mathcal{X}_V \times \mathcal{X}_W} \times \mathcal{X}_J \to [0, 1], \quad (D, x_J) \mapsto P((X_V^{\mathrm{do}(x_J)}, X_W^{\mathrm{do}(x_J)}) \in D)$$

is a Markov kernel $\mathcal{X}_J \dashrightarrow \mathcal{X}_V \times \mathcal{X}_W$. We call it a Markov kernel of $\mathcal{M}$, *and also write it as $P(X_V, X_W \mid \mathrm{do}(X_J))$. Note that it may not be unique.*

If only a subset $O \subseteq V \cup W$ is observed, we also call the marginal Markov kernel $\mathcal{X}_J \dashrightarrow \mathcal{X}_O$, also written as $P_{\mathcal{M}}(X_O \mid \mathrm{do}(X_J))$, an (observed) Markov kernel of $\mathcal{M}$.

Note that if an SCM has multiple solutions, it has multiple Markov kernels, and the notation "$P(X_V, X_W \mid \mathrm{do}(X_J))$" is ambiguous since it does not specify which solution the kernel comes from. Later, we will mostly restrict attention to simple SCMs for which the Markov kernel turns out to be unique. We can construct Markov kernels by pushing the exogenous distribution through a solution function.

Remark 2.8.2. *Given an SCM $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ and a solution function $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$, a Markov kernel of $\mathcal{M}$ is obtained as the push-forward*

$$P(X_V, X_W \mid \mathrm{do}(X_J)) = (g, \mathrm{id}_{\mathcal{X}_W})_* P(X_W \mid X_J)$$

of the Markov kernel

$$P(X_W \mid \mathrm{do}(X_J)) := \bigotimes_{w \in W} P(X_w),$$

where we interpret the exogenous distribution specified by $\mathcal{M}$ as a constant Markov kernel $\mathcal{X}_J \dashrightarrow \mathcal{X}_W$.

Again, not all Markov kernels of an SCM can be obtained in this way in case the SCM admits multiple solution functions (see also Example 5.6.9).

JM: TODO: explain this in terms of a sampling algorithm.

By first intervening on the SCM, constructing a Markov kernel for the intervened SCM, and then conditioning and marginalizing it, the SCM leads to a rich family of Markov kernels that forms a nice playground for causal reasoning, inference, discovery and learning.

3. Graph Theory

3.1. Core Concepts

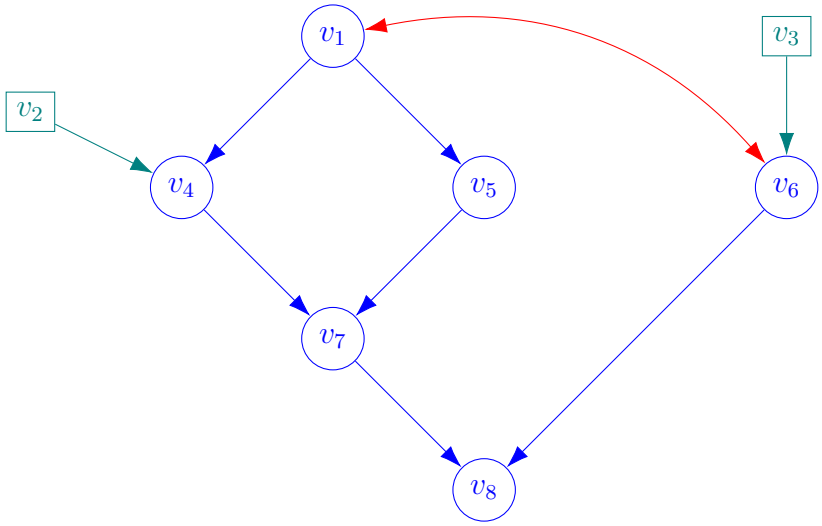


Figure 3: Contextual/Conditional Acyclic Directed Mixed Graph (CADMG).

Definition 3.1.1 (Contextual/conditional directed mixed graphs (CDMG))**.** *A* contextual/conditional directed mixed graph (CDMG) G *- per definition - consists of two (disjoint) sets of vertices/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. $v_1 \rightarrowtail v_2 \in G$ *to mean* $(v_1, v_2) \in E$,
3. $v_1 \leftarrowtail v_2 \in G$ *to mean* $(v_2, v_1) \in E$,
4. $v_1 \leftrightarrow v_2 \in G$ *to mean* $(v_1, v_2) \in L$,
5. $v_1 \nrightarrow v_2 \in G$ *to mean that either* $v_1 \rightarrowtail v_2 \in G$ *or* $v_1 \leftrightarrow v_2 \in G$.

6. $v_1 \hookleftarrow\!\!\ast\, v_2 \in G$ *to mean that either* $v_1 \hookleftarrow v_2 \in G$ *or* $v_1 \leftrightarrow v_2 \in G$.

7. $v_1 \ast\!\!\hookleftarrow\!\!\ast\, v_2 \in G$ *to mean that either* $v_1 \rightarrowtail v_2 \in G$ *or* $v_1 \hookleftarrow v_2 \in G$ *or* $v_1 \leftrightarrow v_2 \in G$.

8. *etc.*

The star on the arrow end here stands for a placeholder to mean: "arrow head or tail".

Remark 3.1.3. *With the notations 3.1.2 the restrictions in definition 3.1.1 mean that the nodes $j \in J$ will not have any arrow heads pointing towards them: $j \hookleftarrow\!\!\ast\, v \notin G$. Nodes $j \in J$ can only point towards nodes $v \in V$: edges $j \rightarrowtail v$ are allowed.*

Definition 3.1.4 (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 sequence of nodes and edges*

$$v = v_0 \ast\!\!\hookleftarrow\!\!\ast\, v_1 \ast\!\!\hookleftarrow\!\!\ast\, \cdots \ast\!\!\hookleftarrow\!\!\ast\, v_{n-1} \ast\!\!\hookleftarrow\!\!\ast\, v_n = w$$

   in G for some $n \geq 0$, i.e. such that for every $k = 1, \ldots, n$ we have that $v_{k-1} \ast\!\!\hookleftarrow\!\!\ast\, v_k \in G$, and with $v_0 = v$ and $v_n = w$.

   In the definition of walk the appearance of the same nodes several times is allowed. Also the trivial walk *consisting of a single node $v_0 \in G$ is allowed as well (if $v = w$).*

2. *A* directed walk *from v to w in G is of the form:*

$$v = v_0 \rightarrowtail v_1 \rightarrowtail \cdots \rightarrowtail v_{n-1} \rightarrowtail v_n = w,$$

   for some $n \geq 0$, where all arrow heads point in the direction of w and there are no arrow heads 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:*

$$v = v_0 \ast\!\!\hookleftarrow\!\!\ast\, v_1 \leftrightarrow \cdots \leftrightarrow v_{n-1} \hookleftarrow\!\!\ast\, v_n = w,$$

   for some $n \geq 0$, where all nodes in between v and w have two arrow heads pointing towards them (aka collider). Note that for $n = 1$ this reads: $v \ast\!\!\hookleftarrow\!\!\ast\, w \in G$.

5. *A walk is called* path *if all occurring nodes are pairwise different.*

6. *A* trek *from v to w in G is a path that is either trivial or of the form:*

$$v = v_0 \leftarrowtail v_1 \leftarrowtail \cdots \leftarrowtail v_{k-1} \leftarrow\!* \; v_k \rightarrowtail \cdots \rightarrowtail v_{n-1} \rightarrowtail v_n = w,$$

with all nodes $v_i \in V$ for $i = 0, \ldots, n$, i.e. where every node has at most one arrow head pointing towards it and the endnodes have an arrow head pointing towards them.

Definition 3.1.5 (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:*

$$\mathrm{Pa}^G(v) := \{w \in G \,|\, w \rightarrowtail v \in G\}.$$

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:*

$$\mathrm{Ch}^G(v) := \{w \in G \,|\, v \rightarrowtail w \in G\}.$$

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:*

$$\mathrm{Sib}^G(v) := \{w \in G \,|\, v \leftrightarrowtail w \in G\}.$$

4. *The set of* ancestors *of v in G:*

$$\mathrm{Anc}^G(v) := \{w \in G \,|\, \exists \text{ directed walk: } w \rightarrowtail \cdots \rightarrowtail v \in G\}.$$

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:*

$$\mathrm{Desc}^G(v) := \{w \in G \,|\, \exists \textit{ directed walk: } v \rightarrowtail \cdots \rightarrowtail w \in G\}\,.$$

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: $A \subseteq \mathrm{Dist}^G(A)$.

9. *The d-*Markov blanket *of v in G:*

$$\mathrm{MBl}_d^G(v) := \{w \in G \,|\, \exists \textit{ collider walk: } v \ast\!\!\rightarrow v_1 \leftrightarrow \cdots \leftrightarrow v_{n-1} \leftarrow\!\!\ast\, w \in G\} \setminus \{v\}.$$

Note: $\big(\mathrm{Pa}^G(v) \cup \mathrm{Ch}^G(v) \cup \mathrm{Pa}^G(\mathrm{Ch}^G(v))\big) \setminus \{v\} \subseteq \mathrm{MBl}_d^G(v)$.

The d-Markov blanket *of* A *in* G*:*

$$\mathrm{MBl}_d^G(A) := \bigcup_{v \in A} \mathrm{MBl}_d^G(v) \setminus A.$$

10. *The* σ-Markov blanket *of* A *in* G*:*

$$\mathrm{MBl}_\sigma^G(A) := \mathrm{MBl}_d^{G^{\mathrm{acy}}}(A),$$

where G^{acy} *is an acyclification of* G*, a notion defined below in Definition 3.4.1.*

Definition 3.1.6 (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 3.1.7. *A Contextual/Conditional Directed Mixed Graph (CDMG)* $G = (J, V, E, L)$ *is called:*

1. *Contextual/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. *Contextual/Conditional Directed Graph (CDG) if* $L = \emptyset$*.*
5. *Directed Graph (DG) if* $J = \emptyset$ *and* $L = \emptyset$*.*
6. *Contextual/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 3.1.8 (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 $J \cup V = \{v_1, \ldots, v_K\}$ *where parents always precede their children.*

Lemma 3.1.9. *A CDMG* $G = (J, V, E, L)$ *is acyclic if and only if it has a topological order.*

Definition 3.1.10 (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:

$$\mathrm{Pred}_\leq^G(v) := \{w \in G \,|\, w < v\} \cup \{v\}.$$

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:*

i.) $J_{\mathrm{do}(W)} := J \cup W$,

ii.) $V_{\mathrm{do}(W)} := V \setminus W$,

iii.) $E_{\mathrm{do}(W)} := E \setminus \{v \rightarrowtail w \,|\, v \in G, w \in W\}$,

iv.) $L_{\mathrm{do}(W)} := L \setminus \{v \dashleftarrow w \,|\, v \in G, w \in W\}$,

where we turn all nodes from W into input nodes and remove all edges with arrow heads towards 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:*

$$\left(G_{\mathrm{do}(W_1)}\right)_{\mathrm{do}(W_2)} = \left(G_{\mathrm{do}(W_2)}\right)_{\mathrm{do}(W_1)} = G_{\mathrm{do}(W_1 \cup W_2)}.$$

3.2.2. Soft Interventions on Graphs

Definition 3.2.1 (Extending CDMGs with soft 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* soft intervention nodes *$I_W = \{I_w \,|\, w \in W\}$ 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.) $J_{\mathrm{do}(I_W)} := J \dot{\cup} \{I_w \,|\, w \in W\}$,

ii.) $V_{\mathrm{do}(I_W)} := V$,

iii.) $E_{\mathrm{do}(I_W)} := E \dot{\cup} \{I_w \rightarrowtail w \,|\, w \in W \setminus J\}$,

iv.) $L_{\mathrm{do}(I_W)} := L$,

where we just add edges $I_w \rightarrowtail w$ for $w \in W \setminus J$, where I_w will represent soft intervention nodes.

Remark 3.2.2. *If G is acyclic then also $G_{\mathrm{do}(I_W)}$ is acyclic and a topological order for for $G_{\mathrm{do}(I_W)}$ is also one for G. Any topological order of G can be extended to one for $G_{\mathrm{do}(I_W)}$ by ordering all the I_w nodes first.*

Lemma 3.2.3 (Hard and soft 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:*

$$\left(G_{\mathrm{do}(I_{W_1})}\right)_{\mathrm{do}(I_{W_2})} = \left(G_{\mathrm{do}(I_{W_2})}\right)_{\mathrm{do}(I_{W_1})} = G_{\mathrm{do}(I_{W_1 \cup W_2})}.$$

We also have:

$$\left(G_{\mathrm{do}(I_{W_1})}\right)_{\mathrm{do}(W_2)} = \left(G_{\mathrm{do}(W_2)}\right)_{\mathrm{do}(I_{W_1})} = G_{\mathrm{do}(I_{W_1}, W_2)}.$$

3.2.3. Marginalization of Graphs

Definition 3.2.1 (Marginalization aka latent projection on CDMGs). *Let $G = (J, V, E, L)$ be a CDMG and $W \subseteq V$ a subset of nodes from V. Then the* marginalization of G w.r.t. W *or the* latent projection of G onto $J \cup V \setminus W$ *is the CDMG: $G^{\setminus W} = (J^{\setminus W}, V^{\setminus W}, E^{\setminus W}, L^{\setminus W})$, where::*

i.) $J^{\setminus W} := J$,

ii.) $V^{\setminus W} := V \setminus W$,

iii.) $E^{\setminus W}$ *consists of all directed edges* $\underline{v} \rightarrowtail \overline{v}$ *with* $\underline{v}, \overline{v} \in J \cup V \setminus W$ *for which there exists a directed walk in G:*

$$\underline{v} \rightarrowtail w_1 \rightarrowtail \cdots \rightarrowtail w_{n-1} \rightarrowtail \overline{v},$$

where all intermediate nodes $w_1, \ldots, w_{n-1} \in W$ (if any).

iv.) $L^{\setminus W}$ *consists of all bi-directed edges* $\underline{v} \leftrightarrowtail \overline{v}$ *with* $\underline{v}, \overline{v} \in V \setminus W$, $\underline{v} \neq \overline{v}$, *for which there exists a trek in G:*

$$\underline{v} \leftarrowtail w_1 \leftarrowtail \cdots \leftarrowtail w_{k-1} \leftarrowtail\!\!* \, w_k \rightarrowtail \cdots \rightarrowtail w_{n-1} \rightarrowtail \overline{v},$$

where all intermediate nodes $w_1, \ldots, w_{n-1} \in W$ (if any).

Remark 3.2.2 (Marginalization preserves ancestral relations and acyclicity). *1. For $v_1, v_2 \in G$ with $v_1, v_2 \notin W$ we have the equivalence:*

$$v_1 \in \mathrm{Anc}^{G}(v_2) \quad \Longleftrightarrow \quad v_1 \in \mathrm{Anc}^{G^{\setminus W}}(v_2).$$

2. If the CDMG G is acyclic then also $G^{\setminus W}$ and a topological order of G induces a topological order on $G^{\setminus W}$ (by just ignoring the nodes from W).

Lemma 3.2.3 (Marginalizations commute). *Let $G = (J, V, E, L)$ be a CDMG and $W_1, W_2 \subseteq V$ two disjoint subsets of nodes from V. 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 3.2.4 (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:*

$$\left(G_{\mathrm{do}(W_1)}\right)^{\setminus W_2} = \left(G^{\setminus W_2}\right)_{\mathrm{do}(W_1)}.$$

So we can without ambiguity also write: $G^{\setminus W_2}_{\mathrm{do}(W_1)}$.
A similar statement holds for soft interventions.

3.3. d-Separation and Sigma-Separation

Definition 3.3.1 (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[9] *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:

$$\begin{aligned}
&\textit{left chain:} && v_{k-1} \leftarrowtail v_k \ast\!\!\leftarrow v_{k+1} && \textit{with} && v_k \in C,\\
&\textit{right chain:} && v_{k-1} \ast\!\!\rightarrow v_k \rightarrowtail v_{k+1} && \textit{with} && v_k \in C,\\
&\textit{fork:} && v_{k-1} \leftarrowtail v_k \rightarrowtail v_{k+1} && \textit{with} && v_k \in C,\\
&\textit{collider:} && v_{k-1} \ast\!\!\rightarrow v_k \ast\!\!\leftarrow v_{k+1} && \textit{with} && v_k \notin C.
\end{aligned}$$

   If we consider end nodes, left/right chain and fork as non-colliders *then we can simply say:*
   π is d-blocked by C if it either contains a non-collider in C or a collider not in C.

2. *We say that the walk π is* C-d-open *if it is not C-d-blocked, i.e. if all non-colliders are not in C and all collider are in C.*

Definition 3.3.2 (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:*

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

1. *We say that the walk π is* C-σ-blocked *or* σ-blocked by C *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:

$$\begin{aligned}
&\textit{left chain:} && v_{k-1} \leftarrowtail v_k \ast\!\!\leftarrow v_{k+1} && \textit{with} && v_k \in C \wedge v_k \notin \mathrm{Sc}^G(v_{k-1}),\\
&\textit{right chain:} && v_{k-1} \ast\!\!\rightarrow v_k \rightarrowtail v_{k+1} && \textit{with} && v_k \in C \wedge v_k \notin \mathrm{Sc}^G(v_{k+1}),\\
&\textit{fork:} && v_{k-1} \leftarrowtail v_k \rightarrowtail v_{k+1} && \textit{with} && v_k \in C \wedge v_k \notin \mathrm{Sc}^G(v_{k-1}) \cap \mathrm{Sc}^G(v_{k+1}),\\
&\textit{collider:} && v_{k-1} \ast\!\!\rightarrow v_k \ast\!\!\leftarrow v_{k+1} && \textit{with} && v_k \notin C.
\end{aligned}$$

2. *We say that the walk π is* C-σ-open *if it is not C-σ-blocked.*

Remark 3.3.3. *In words: the walk π is* C-σ-open *if and only if:*

1. *both endpoints of π are not in C, and*

[9] 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] to emphasize the use of the bi-directed edges. But since the notion of d-separation can be found in standard books and m-separation is the natural extension of d-separation we will call it d-separation again.

2. *all colliders on π are in C, and*

3. *all non-endpoint non-colliders on π are not in C or only point to neighboring nodes in the same strongly connected component.*

Definition 3.3.4 (d-separation/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 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!) is d-blocked by C.

Otherwise we write:

$$A \overset{d}{\underset{G}{\not\perp}} B \,|\, C.$$

2. 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.

Otherwise we write:

$$A \overset{\sigma}{\underset{G}{\not\perp}} B \,|\, C.$$

Special case:

$$A \overset{d}{\underset{G}{\perp}} B \qquad :\Longleftrightarrow \qquad A \overset{d}{\underset{G}{\perp}} B|\emptyset.$$

Remark 3.3.5. *If G is acyclic then $\mathrm{Sc}^G(v) = \{v\}$ for all $v \in G$ and the additional conditions in "σ-blocked" for the non-colliders are of the form $v_k \notin \mathrm{Sc}^G(v_{k\pm1}) = \{v_{k\pm1}\}$. These are then automatically satisfied, because the node v_k is a parent of the relevant other node $v_{k\pm1}$ and thus not equal to it. So in the acyclic case d-separation and σ-separation are equivalent. It turns out that in the non-acyclic case σ-separation is the better concept (and as said above it also captures the acyclic case equivalently well), see [FM17, FM18, FM20]. 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.*

Lemma 3.3.6 (d-separation/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 subset of nodes such that:*

$$D \cap (A \cup B \cup C) = \emptyset.$$

Then we have the equivalence:

$$A \overset{d}{\underset{G}{\perp}} B \,|\, C \qquad \Longleftrightarrow \qquad A \overset{d}{\underset{G^{\setminus D}}{\perp}} B \,|\, C.$$

We also have:

$$A \overset{\sigma}{\underset{G}{\perp}} B \,|\, C \qquad \Longleftrightarrow \qquad A \overset{\sigma}{\underset{G^{\setminus D}}{\perp}} B \,|\, C.$$

3.3.1. (Alternative definition of σ-blocking)

PF: Alernative definition of σ-blockking moved to Graph Theory in the begining.

Revisit Definition 3.3.2 of σ-blocking. In the literature, one also encounters a slightly different notion with the same name. In this optional section, we will clarify the relationship between the two. For the rest of the course, we will keep using Definition 3.3.2.

Definition 3.3.1. *Let $G = \langle J, V, E, L\rangle$ be a CDMG, $C \subseteq J \cup V$ a subset of nodes and π a walk (or path) in G:*

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

1. *We say that the walk (or path) π is C-σ'-*blocked *or σ'-*blocked *by C 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:

$$\begin{array}{rlll}
\textit{left chain:} & v_{k-1} \leftarrow v_k \leftarrow\!\ast\, v_{k+1} & \textit{with} & v_k \in C \wedge v_k \notin \mathrm{Sc}^G(v_{k-1}),\\
\textit{right chain:} & v_{k-1} \ast\!\!\rightarrow v_k \rightarrow v_{k+1} & \textit{with} & v_k \in C \wedge v_k \notin \mathrm{Sc}^G(v_{k+1}),\\
\textit{fork:} & v_{k-1} \leftarrow v_k \rightarrow v_{k+1} & \textit{with} & v_k \in C \wedge v_k \notin \mathrm{Sc}^G(v_{k-1}) \cap \mathrm{Sc}^G(v_{k+1}),\\
\textit{collider:} & v_{k-1} \ast\!\!\rightarrow v_k \leftarrow\!\ast\, v_{k+1} & \textit{with} & v_k \notin \mathrm{Anc}^G(C).
\end{array}$$

2. *We say that the walk (or path) π is C-σ'-*open *if it is not C-σ'-blocked.*

The only difference is in the condition for a collider node v_k: to σ-block a walk one needs $v_k \notin C$, whereas to σ'-block a walk one needs $v_k \notin \mathrm{Anc}^G(C)$.

The following lemma will be convenient to relate the notions.

Lemma 3.3.2. *Let $G = \langle V, J, E, L\rangle$ be a CDMG, $C \subseteq V \cup J$ and $\pi = (v_0, \ldots, v_n)$ be a C-σ'-open walk in G. Suppose $v_i \in \mathrm{Sc}^G(v_j)$ with $i < j$. We can then replace the subwalk $v_i, \ldots, v_j$ of π by*

(i) a shortest directed path $v_i \rightarrow \cdots \rightarrow v_j$ in G if $j = n$ or if $v_j \rightarrow v_{j+1}$ on π, or

(ii) a shortest directed path $v_i \leftarrow \cdots \leftarrow v_j$ in G otherwise,

that is entirely within $\mathrm{Sc}^G(v_i)$ and such that the modified walk π' is still C-σ'-open.

Proof. π' cannot become C-σ'-blocked by one of the initial nodes $v_0 \dots v_{i-1}$ or 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 do not C-σ'-block π by assumption. Furthermore, π' cannot become C-σ'-blocked through one of the nodes strictly between v_i and v_j on π' (if there are any), since these nodes are all non-endpoint non-colliders that only point to nodes in the same strongly connected component on π'. Because π is C-σ'-open, $v_j \notin C$ if $j = n$ or if $v_j \to i_{j+1}$ on π. This holds in particular in case (i). Similarly, $v_i \notin C$ if $i = 0$ or $v_{i-1} \leftarrow v_i$ on π.

In case (i), π' is not C-σ'-blocked by v_j because v_j is a non-collider on π' but $v_j \notin C$. Also v_i does not C-σ'-block π'. Indeed, assume $v_i \neq v_j$ (otherwise there is nothing to prove). If $i = 0$, or if $i > 0$ and $v_{i-1} \leftarrow v_i$ on π', then the same holds for π and hence $v_i \notin C$; v_i is then a non-collider on π', but $v_i \notin C$. If $i > 0$ and $v_{i-1} \ast\!\!\to v_i$ on π' then v_i is a non-endpoint non-collider on π' that does not point to a node in another strongly connected component.

Now consider case (ii). If $i = 0$ or $v_{i-1} \leftarrow v_i$ on π' then this case is analogous to case (i). So assume $i > 0$ and $v_{i-1} \ast\!\!\to v_i$ on π'. v_i must be a collider on π' (whether $v_i = v_j$ or not). Then on the subwalk of π between v_i and v_j there must be a directed walk from v_i to a collider that is in $\mathrm{Anc}^G(C)$, which implies that $v_i \in \mathrm{Anc}^G(C)$. Hence, v_i will not C-σ'-block π'. Also v_j cannot C-σ'-block π'. Indeed, assume $v_i \neq v_j$ (otherwise there is nothing to prove). Since $v_j \leftarrow\!\!\ast\, v_{j+1}$ on π', v_j is a non-endpoint non-collider on π' that does not point to a node in another strongly connected component. □

With the help of this lemma, we can prove that Definition 3.3.1 indeed serves as an alternative to Definition 3.3.2.

Proposition 3.3.3. *Let $G = \langle J, V, E, L \rangle$ be a CDMG. For $C \subseteq J \cup V$, and $i, j \in J \cup V$, the following are equivalent:*

1. *there exists a C-σ-open walk between i and j in G;*
2. *there exists a C-σ'-open walk between i and j in G;*
3. *there exists a C-σ'-open path between i and j in G.*

Proof. 1 $\implies$ 2: Suppose there exists a C-σ-open walk between i and j. All colliders are in C, and hence in $\mathrm{Anc}^G(C)$. So this same walk is also C-σ'-open.

2 $\implies$ 1: Suppose there exists a C-σ'-open walk between i and j. Then all colliders are in $\mathrm{Anc}^G(C)$. If a collider is not in C, we can extend the walk by replacing this collider $\ast\!\!\to v \leftarrow\!\!\ast$ by the walk $\ast\!\!\to v \to \cdots \to c \leftarrow \cdots \leftarrow v \leftarrow\!\!\ast$ consisting of a directed walk starting from v towards the first node c in C it encounters, and from there tracing back the same walk in opposite direction to v. All nodes on this walk except for c (v included) are non-colliders not in C. Extending this walk in such a way for each collider not in C, we obtain an extended walk that is C-σ-open.

2 $\implies$ 3: Let $\pi = (v_0, \dots, v_n)$ be a C-σ'-open walk between nodes v_0 and v_0 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 3.3.2 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 v_0 and v_n.

3 $\implies$ 2: trivial. □

Hence, it does not matter which of the two notions we use in the definition of σ-separation. When checking σ-separation in a graph "by hand", we usually use the third formulation, since there are only a finite number of paths to check. 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).

3.4. Acyclifications

PF: Graph part of acyclification moved to Graph Theory in the begining.

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 for IL-CBNs to cyclic SCMs. The construction was first proposed in the context of CBNs by [Spi94, Spi95].

Definition 3.4.1. *Given a CDMG $G = \langle J, V, E, L\rangle$, we call a CADMG $G' = \langle J, V, E', L'\rangle$ an* acyclification of G *if*

(i) G' *is acyclic;*

(ii) G' *has the same input nodes J and output nodes V as G;*

(iii) for any pair of nodes $\{i, j\}$ such that $i \notin \mathrm{Sc}^G(j)$:

a) $i \to j \in E'$ *iff there exists a node $j' \in \mathrm{Sc}^G(j)$ such that $i \to j' \in E$;*

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 any pair of distinct nodes $\{i, j\}$ such that $i \in \mathrm{Sc}^G(j)$: $i \to j \in \mathcal{E}'$ or $i \leftarrow j \in \mathcal{E}'$ or $i \leftrightarrow j \in \mathcal{F}'$.

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 3.4.2. *Let $G = \langle J, V, E, L\rangle$ be a CDMG and $G' = \langle J, V, E', L'\rangle$ an acyclification of G. Then for $A, B, C \subseteq V \cup J$ (not necessarily disjoint) subsets of nodes,*

$$A \overset{\sigma}{\underset{G}{\perp}} B \,|\, C \iff A \overset{\sigma}{\underset{G'}{\perp}} B \,|\, C \iff A \overset{d}{\underset{G'}{\perp}} B \,|\, C$$

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 $\pi = (v_0, \ldots, v_n)$ 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 $(v_i, \ldots, v_j)$ 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) $*\!\!\rightarrow v_i \cdots v_j \leftarrow\!\!*$: 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. We can replace this subwalk by $*\!\!\rightarrow w \leftarrow\!\!*$ in G' such that w becomes a collider in C.

(ii) $(\leftarrow)v_i \cdots v_j \leftarrow\!\!*$:[10] 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 $*\!\!\rightarrow v_i \rightarrow \cdots \rightarrow w \leftarrow \cdots \leftarrow v_j \leftarrow\!\!*$ 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.

[10] We put parentheses around the first directed edge to indicate that this case also applies if v_i is an endnode, i.e., if $i = 0$.

(ii) $(\leftarrow)v_i \cdots v_j \leftrightarrow *$: 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 $(\leftarrow)v_i \leftarrow \cdots \leftarrow v_j \leftrightarrow *$ in G with all nodes in $\mathrm{Sc}^G(v_i)$. Note that the modified walk remains C-σ-open.

(iii) $*\leftrightarrow v_i \cdots v_j(\rightarrow)$: analogous to the previous case, we can replace it by $*\leftrightarrow v_i \rightarrow \cdots \rightarrow v_j(\rightarrow)$ 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 $(\leftarrow)v_i \rightarrow \cdots \rightarrow v_j(\rightarrow)$ in G with all nodes in $\mathrm{Sc}^G(v_i)$. The modified walk remains C-σ-open.

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 $i \rightarrow j$ on π' with $j \notin \mathrm{Sc}^G(i)$, there must be a $j' \in \mathrm{Sc}^G(j)$ such that $i \rightarrow j'$ is present in G, and hence there must be a directed path $j' \rightarrow \cdots \rightarrow j$ entirely in $\mathrm{Sc}^G(j)$ such that we can use $i \rightarrow j' \rightarrow \cdots \rightarrow j$ as replacement in G of the edge $i \rightarrow j$. 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 $i \leftarrow \cdots \leftarrow i' \leftrightarrow j' \rightarrow \cdots \rightarrow j$ in G, where $i \leftarrow \cdots \leftarrow i'$ is entirely in $\mathrm{Sc}^G(i)$ and $j' \rightarrow \cdots \rightarrow j$ 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. □

PF: show that acyclification is acyclic

3.5. Separoid Axioms for d-/Sigma-Separation

Definition/Theorem 3.5.1 ((Asymmetric) separoid axioms for d-separation/σ-separation)**.** *Let $G = (J, V, E, L)$ be a CDMG and $A, B, C, D \subseteq J \cup V$ subset of nodes. Then the ternary relations $\perp = \perp_G^d$ and $\perp = \perp_G^\sigma$ satisfy the following rules:*

a) Extended Left Redundancy:

$$D \subseteq A \implies D \perp B \,|\, A.$$

b) J-Restricted Right Redundancy:

$A \perp \emptyset \,|\, C \cup J$ *always holds.*

c) J-Inverted Right Decomposition:

$$A \perp B \,|\, C \implies A \perp J \cup B \,|\, C.$$

d) Left Decomposition:

$$A \cup D \perp B \,|\, C \implies D \perp B \,|\, C.$$

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:

$$(A \perp B \cup D \,|\, C) \iff (A \perp B \,|\, D \cup C) \wedge (A \perp D \,|\, C),$$

$$(A \cup D \perp B \,|\, C) \iff (A \perp B \,|\, D \cup C) \wedge (D \perp B \,|\, C).$$

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 3.5.2. *Let the assumptions be like in 3.5.1. We also have the following rules:*

l) Left Composition:

$$(A \perp B \,|\, C) \wedge (D \perp B \,|\, C) \implies A \cup D \perp B \,|\, C.$$

m) Right Composition:

$$(A \perp B \,|\, C) \wedge (A \perp D \,|\, C) \implies A \perp B \cup D \,|\, C.$$

n) Left Intersection: If $A \cap D = \emptyset$ then:

$$(A \perp B \,|\, D \cup C) \wedge (D \perp B \,|\, A \cup C) \implies A \cup D \perp B \,|\, C.$$

m) 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 d-/Sigma-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.
Since σ-separation can be expressed as d-separation in an acyclification G' of G (see Definition 3.4.1) by Proposition 3.4.2

$$A \overset{\sigma}{\underset{G}{\perp}} B \,|\, C \qquad \iff \qquad A \overset{d}{\underset{G'}{\perp}} B \,|\, C,$$

we can w.l.o.g. in the proof assume that we only use d-separation.

Recall that we say that 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!) is d-blocked by C.
Again a walk π is *d-blocked by* C if it either contains a non-collider (i.e. either an end node, fork, left/right chain) in C or a collider not in C.

We abbreviate the ternary relations in the following: $\perp := \overset{d}{\underset{G}{\perp}}$.

Lemma 3.5.3 (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 d-blocked by A. □

Lemma 3.5.4 (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 d-blocked by A. □

Lemma 3.5.5 (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 d-blocked by C. □

Lemma 3.5.6 (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 d-blocked by C. □

Lemma 3.5.7 (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 d-blocked by C. □

Lemma 3.5.8 (Left Weak Union)**.**

$$A \cup D \perp B \,|\, C \implies A \perp B \,|\, D \cup C.$$

Proof. Lets assume the contrary: $A \not\perp B \,|\, D \cup C$. Then there exists a shortest $(D \cup C)$-d-open walk π from a node v in A to a node w in $J \cup B$ in G. Then every collider of π is in $D \cup C$ and every 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 lies in C. This implies that π is C-d-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 from $w \in J \cup B$ 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. So $\tilde{\pi}$ is C-d-open walk from $A \cup D$ to $J \cup B$. This contradicts the assumption: $A \cup D \perp B \,|\, C$. □

Lemma 3.5.9 (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 3.5.8, but this time get a contradiction with: $A \perp B \cup D \,|\, C$. 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. So $\tilde{\pi}$ is C-d-open walk from A to $J \cup B \cup D$. This contradicts the assumption: $A \perp B \cup D \,|\, C$. □

Lemma 3.5.10 (Left Contraction)**.**

$$(A \perp B \,|\, D \cup C) \wedge (D \perp B \,|\, C) \implies A \cup D \perp B \,|\, C.$$

Proof. Lets assume the contrary: $A \cup D \not\perp B \,|\, C$. Then there exists a shortest C-d-open walk π from a node v in $A \cup D$ to a node w in $J \cup B$ in G. So every collider of π lies in C and every non-collider lies not in C and v is the only node of π that lies in $(A \cup D) \setminus C$ (otherwise π could be shortend).
Also v cannot lie in $D \setminus C$ as it would contradict the assumption: $D \perp B \,|\, C$. Thus $v \in A \setminus C$ and π is a walk from A to $J \cup B$ whose colliders all lie in $C \subseteq D \cup C$ and all non-collider outside of $D \cup C$. But this contradicts the other assumption: $A \perp B \,|\, D \cup C$. □

Lemma 3.5.11 (Right Contraction)**.**

$$(A \perp B \,|\, D \cup C) \wedge (A \perp D \,|\, C) \implies A \perp B \cup D \,|\, C.$$

Proof. Lets assume the contrary: $A \not\perp B \cup D \,|\, C$. Then there exists a shortest C-d-open walk π from a node v in A to a node w in $J \cup B \cup D$ in G. So every collider of π lies in C and every non-collider outside C and w is the only node of π that lies in $(J \cup B \cup D) \setminus C$ (otherwise π could be shortend).
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 lie in $C \subseteq D \cup C$ and all non-colliders outside of $D \cup C$. But this contradicts the other assumption: $A \perp B \,|\, D \cup C$. □

Lemma 3.5.12 (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 3.5.11, only the first contradiction is with: $D \perp A \,|\, C$. □

Lemma 3.5.13 (Flipped Left Cross Contraction)**.**

$$(A \perp B \,|\, D \cup C) \wedge (B \perp D \,|\, C) \implies B \perp A \cup D \,|\, C.$$

Proof. Lets assume the contrary: $B \not\perp A \cup D \,|\, C$. Then there exists a shortest C-d-open walk π from a node v in B to a node w in $J \cup A \cup D$ in G. So every collider of π lies in C and every non-collider outside C and w is the only node of π that lies in $(J \cup A \cup D) \setminus C$ (otherwise π could be shortend).
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 all lie in $C \subseteq D \cup C$ and all non-colliders outside of $D \cup C$. But this contradicts the other assumption: $A \perp B \,|\, D \cup C$. □

Lemma 3.5.14 (J-Restricted Symmetry)**.**

$$A \perp B \,|\, C \cup J \implies B \perp A \,|\, C \cup J.$$

Proof. This follows from Flipped Left Cross Contraction 3.5.13 with $D = \emptyset$ and $C \cup J$ in place of C together with J-Restricted Right Redundancy 3.5.4. □

Lemma 3.5.15 (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 d-blocked by C by assumption: $A \perp B \,|\, C$. If $v \in D$ then π is d-blocked by C by assumption: $D \perp B \,|\, C$. □

Lemma 3.5.16 (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 d-blocked by C by assumption: $A \perp B \,|\, C$. If $w \in J \cup D$ then π is d-blocked by C by assumption: $A \perp D \,|\, C$. □

Lemma 3.5.17 (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. Lets assume the contrary: $A \cup D \not\perp B \,|\, C$. Then there exists a shortest C-d-open walk π from a node v in $A \cup D$ to a node w in $J \cup B$ in G. So every collider of π lies in C and every non-collider outside C and v is the only node of π that lies in $(A \cup D) \setminus C$ (otherwise π could be shortend).
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 lie in $C \subseteq D \cup C$ and all non-colliders outside of $(D \setminus C) \cup C = D \cup C$. This contradicts the assumption: $A \perp B \,|\, D \cup C$.
If $v \in D$ then similarly we get a contradiction: $D \perp B \,|\, A \cup C$ □

Lemma 3.5.18 (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. Lets assume the contrary: $A \not\perp B \cup D \,|\, C$. Then there exists a shortest C-d-open walk π from a node v in A to a node w in $J \cup B \cup D$ in G. So every collider of π lies in C and every non-collider outside C and w is the only node of π that lies in $(J \cup B \cup D) \setminus C$ (otherwise π could be shortend).
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 lies in $C \subseteq B \cup C$ and all non-colliders are outside of $(B \setminus C) \cup C = B \cup C$. So π is a $(B \cup C)$-d-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 lies in $C \subseteq D \cup C$ and all non-colliders are outside of $D \cup C$. So π is a $(D \cup C)$-d-open walk from A to $J \cup D$. 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. □

4. Causal Bayesian Networks

4.1. Core Concepts

Definition 4.1.1 (Causal Bayesian network with input variables)**.** *A* causal Bayesian network with input variables (I-CBN) - *by definition - consists of:*

a.) a contextual/conditional directed acyclic graph (CDAG): $G = (J, V, E)$ (with finite vertex sets),

b.) a standard measurable space $\mathcal{X}_v$ for every $v \in J \cup V$,

c.) a Markov kernel: $P_v\left(X_v | \operatorname{do}\left(X_{\mathrm{Pa}^G(v)}\right)\right)$:

$$\begin{array}{rcl} \mathcal{X}_{\mathrm{Pa}^G(v)} & \dashrightarrow & \mathcal{X}_v, \\ (A, x_{\mathrm{Pa}^G(v)}) & \mapsto & P_v\left(X_v \in A | \operatorname{do}\left(X_{\mathrm{Pa}^G(v)} = x_{\mathrm{Pa}^G(v)}\right)\right), \end{array}$$

for every $v \in V$, where we write for $D \subseteq J \cup V$:

$$\begin{array}{ll} \mathcal{X}_D := \prod_{v \in D} \mathcal{X}_v, & \mathcal{X}_\emptyset := * = \{*\}, \\ X_D := (X_v)_{v \in D}, & X_\emptyset := *, \\ x_D := (x_v)_{v \in D}, & x_\emptyset := *. \end{array}$$

Definition 4.1.2 (The joint Markov kernel of a causal Bayesian network with input variables)**.** *Consider a causal Bayesian network with input variables with CDAG $G = (J, V, E)$ and Markov kernels $P_v\left(X_v | \operatorname{do}\left(X_{\mathrm{Pa}^G(v)}\right)\right)$ for $v \in V$. For a fixed topological ordering $<$ of G we then define the joint Markov kernel of the I-CBN:*

$$\mathcal{X}_J \dashrightarrow \mathcal{X}_V$$

as follows:

$$P(X_V | \operatorname{do}(X_J)) := \bigotimes_{v \in V, >} P_v\left(X_v | \operatorname{do}\left(X_{\mathrm{Pa}^G(v)}\right)\right),$$

where the nodes v run through V in reverse ordering of $<$, i.e. all parents are on the right of all their children.

Exercise 4.1.3. *Show that the definition of the joint Markov kernel of an I-CBN is independent of the topological ordering.*

Notation 4.1.4. *By abuse of notation, we will refer to the tuple:*

$$(G, P(X_V | \operatorname{do}(X_J)))$$

as the I-CBN, keeping the single Markov kernels P_v and the spaces $\mathcal{X}_v$ implicit.

Remark 4.1.5 (Marginalization and conditioning)**.** *Let $P(X_V|\operatorname{do}(X_J))$ be the joint Markov kernel of a I-CBN. We can extend it to a joint Markov kernel including X_J:*

$$P(X_V, X_J|\operatorname{do}(X_J)) = P(X_V|\operatorname{do}(X_J)) \otimes \delta(X_J|X_J).$$

For any $A, B \subseteq J \cup V$ we then also have the marginal conditional *Markov kernel:*

$$P(X_A|X_B, \operatorname{do}(X_J)),$$

which exists by theorem 2.4.1 due to the use of standard measurable spaces and is unique up to a $P(X_B|\operatorname{do}(X_J))$-null set.
Furthermore, if $C \subseteq J$ and we have:

$$X_A \underset{P(X_V|\operatorname{do}(X_J))}{\perp\!\!\!\perp} X_J \,|\, X_C,$$

then we also have a Markov kernel:

$$P(X_A|\operatorname{do}(X_C))$$

that fits into the equation:

$$P(X_A, X_C|\operatorname{do}(X_J)) = P(X_A|\operatorname{do}(X_C)) \otimes P(X_C|\operatorname{do}(X_J)).$$

Note that this is unique up to a $P(X_C|\operatorname{do}(X_J))$-null set and that $P(X_C|\operatorname{do}(X_J)) = \delta(X_C|X_J)$, and thus $P(X_A|\operatorname{do}(X_C))$ is unique (not just up to null sets). In other words, $P(X_A|\operatorname{do}(X_J))$ is only dependent on the arguments from X_C.

Definition 4.1.6 (Causal Bayesian network with input variables and latent variables)**.** *A* causal Bayesian network with input variables and latent variables *or* latent variable causal Bayesian network with input variables (IL-CBN) - *per definition - consists of:*

a.) an I-CBN with:

a) CDAG: $G^+ = (J, V^+, E^+)$, and:

b) Markov kernel:

$$P\left(X_{V^+}|\operatorname{do}\left(X_J\right)\right) = \bigotimes_{v \in V^+} P_v\left(X_v|\operatorname{do}\left(X_{\operatorname{Pa}^{G^+}(v)}\right)\right),$$

and:

b.) a decomposition of the set of nodes V^+:

$$V^+ = V \dot\cup U,$$

into disjoint sets of observed nodes V *and* unobserved nodes U.

By abuse of notation, we refer to it as the tuple:

$$M = \left(G^+ = (J, V, U, E^+),\ P\left(X_V, X_U | \operatorname{do}(X_J)\right)\right).$$

We refer to the marginal:

$$P\left(X_V | \operatorname{do}(X_J)\right)$$

as the observational Markov kernel. *We call the marginalized CADMG:*

$$G := (J, V, E, L) := (G^+)^{\backslash U}$$

the (induced) observational CADMG.

Remark 4.1.7 (Causal Bayesian networks)**.** *In the following if we say* causal Bayesian network (CBN) *we will mean causal Bayesian network with input variables and with latent variables (IL-CBN).*

4.2. Global Markov Property

Theorem 4.2.1 (Global Markov property for causal Bayesian networks)**.** *Consider a causal Bayesian network (with input and latent variables) with observational CADMG $G = (J, V, E, L)$ and observational 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:*

$$A \overset{d}{\underset{G}{\perp}} B \,|\, C \qquad \Longrightarrow \qquad X_A \underset{P(X_V | \operatorname{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C.$$

If one wants to make the implicit dependence on J more explicit one can equivalently also write:

$$A \overset{d}{\underset{G}{\perp}} J \cup B \,|\, C \qquad \Longrightarrow \qquad X_A \underset{P(X_V | \operatorname{do}(X_J))}{\perp\!\!\!\perp} X_J, X_B \,|\, X_C.$$

Notation 4.2.2. *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:

$$P(X_A | X_{C \cap V}, \operatorname{do}(X_J))$$

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 2.5.11 that we can, 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.
In these suggestive notations we can state the global Markov property 4.2.1 as:

$$\begin{aligned} & A \overset{d}{\underset{G}{\perp}} B \,|\, C \\ \implies \quad & P(X_A|X_B, X_C, \mathrm{do}(X_J)) \quad = P(X_A|\cancel{X_B}, X_{C\cap V}, \mathrm{do}(X_{C\cap J}), \cancel{\mathrm{do}(X_J)}), \quad \text{a.s.} \\ & \qquad\qquad\qquad\qquad\quad = P(X_A|X_{C\cap V}, \mathrm{do}(X_{C\cap J}), \cancel{\mathrm{do}(X_J)}) \quad \text{a.s.} \\ & \qquad\qquad\qquad\qquad\quad = P(X_A|\cancel{\mathrm{do}(X_J)}, X_{C\cap V}, \mathrm{do}(X_{C\cap J})) \quad \text{a.s.} \end{aligned}$$

Proofs - Global Markov Property

The proof of the global Markov property follows similar arguments as used in [LDLL90, Ver93, Ric03, FM17, FM18, RERS17], 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 4.2.3 (Global Markov property for causal Bayesian networks)**.** *Consider a causal Bayesian network (with input and latent variables) with observational CADMG $G = (J, V, E, L)$ and observational 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:*

$$A \overset{d}{\underset{G}{\perp}} B \,|\, C \qquad \Longrightarrow \qquad X_A \underset{P(X_V|\,\mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C.$$

If one wants to make the implicit dependence on J more explicit one can equivalently also write:

$$A \overset{d}{\underset{G}{\perp}} J \cup B \,|\, C \qquad \Longrightarrow \qquad X_A \underset{P(X_V|\,\mathrm{do}(X_J))}{\perp\!\!\!\perp} X_J, X_B \,|\, X_C.$$

Proof. Because d-separation is preserved under marginalization:

$$A \overset{d}{\underset{G}{\perp}} B \,|\, C \quad \Longleftrightarrow \quad A \overset{d}{\underset{G^+}{\perp}} B \,|\, C,$$

we can directly assume that we work with the causal Bayesian networks 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:

$$A \overset{d}{\underset{G}{\perp}} J \cup B \,|\, C,$$

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:

$$P(X_A, X_B, X_C | \operatorname{do}(X_J)) = \underbrace{\bigotimes_{w \in A} \delta(X_w | X_w)}_{=:Q(X_A|X_C)} \otimes \underbrace{\bigotimes_{w \in B} \delta(X_w | X_w) \otimes \bigotimes_{w \in C} \delta(X_w | X_w)}_{=P(X_B, X_C | \operatorname{do}(X_J))}.$$

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:

$$X_A \underset{P(X_V | \operatorname{do}(X_J))}{\perp\!\!\!\perp} X_B | X_C.$$

(IND): Induction assumption: The global Markov property holds for all causal Bayesian networks (with input variables, but without latent variables and bi-directed edges) with $\#V < n$ (and arbitrary J).

1.) Now assume: $\#V = n > 0$ and $A \perp_G^d J \cup B \,|\, C$.

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 $\operatorname{Ch}^G(v) = \emptyset$, 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:

$$P(X_V | \operatorname{do}(X_J)) = P_v(X_v | \operatorname{do}(X_{\operatorname{Pa}^G(v)})) \otimes \underbrace{\bigotimes_{w \in \operatorname{Pred}^G_<(v) \setminus J} P_w(X_w | \operatorname{do}(X_{\operatorname{Pa}^G(w)}))}_{P(X_{\operatorname{Pred}^G_<(v) \setminus J} | \operatorname{do}(X_J))}.$$

This factorization implies that we already have the conditional independence:

$$X_v \underset{P(X_V | \operatorname{do}(X_J))}{\perp\!\!\!\perp} X_{\operatorname{Pred}^G_<(v)} \,|\, X_D,$$

where we put $D := \operatorname{Pa}^G(v)$.

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:

$$A \overset{d}{\underset{G}{\perp}} J \cup B \,|\, C,$$

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:

$$A \overset{d}{\underset{G}{\perp}} J \cup B \,|\, C \quad \Longleftrightarrow \quad A \overset{d}{\underset{G^{\setminus v}}{\perp}} J \cup B \,|\, C.$$

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:

$$\begin{aligned} A \overset{d}{\underset{G}{\perp}} J \cup B \,|\, C &\xRightarrow{\text{Right Decomposition}} && A \overset{d}{\underset{G}{\perp}} J \cup B' \,|\, C \\ &\xRightarrow{\text{Left Decomposition}} && A' \overset{d}{\underset{G}{\perp}} J \cup B' \,|\, C \\ &\xRightarrow{\text{marginalization, } v \notin A' \cup J \cup B' \cup C} && A' \overset{d}{\underset{G^{\setminus\{v\}}}{\perp}} J \cup B' \,|\, C \\ &\xRightarrow{\text{induction (IND)}} && X_{A'} \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{B'} \,|\, X_C. \qquad (\#1) \end{aligned}$$

On the other hand we have with $D = \mathrm{Pa}^G(v)$:

$$
\begin{aligned}
A \overset{d}{\underset{G}{\perp}} J \cup B \,|\, C \quad &\xRightarrow{\text{Right Decomposition, } B' \subseteq B} && A \overset{d}{\underset{G}{\perp}} J \cup B' \,|\, C \\
&\xRightarrow{\text{Left Weak Union, } A = A' \dot\cup (A \cap C) \dot\cup \{v\}} && \{v\} \overset{d}{\underset{G}{\perp}} J \cup B' \,|\, A' \dot\cup C. \\
&\xRightarrow{(*), \text{ see below}} && D \overset{d}{\underset{G}{\perp}} J \cup B' \,|\, A' \dot\cup C \\
&\xRightarrow{\text{marginalization, } v \notin D \cup J \cup B' \cup A' \cup C} && D \overset{d}{\underset{G^{\setminus \{v\}}}{\perp}} J \cup B' \,|\, A' \dot\cup C \\
&\xRightarrow{\text{induction (IND)}} && X_D \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{B'} \,|\, X_{A' \dot\cup C} \\
&\xRightarrow{A' \dot\cup C} && X_D \underset{P(X_V \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{B'} \,|\, X_{A'}, X_C. \qquad (\#2)
\end{aligned}
$$

(*) holds since every $(A' \dot\cup C)$-open walk $w \ast\!\!\leftarrow\!\!\ast \cdots$ 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 $v \leftarrow\!\!\leftarrow w \ast\!\!\leftarrow\!\!\ast \cdots$, 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:

$$
\begin{aligned}
& X_{A'}, X_v, X_C \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \mid X_{A'}, X_v, X_C \\
\xRightarrow{\text{Left Contraction, (\#3)}} \quad & X_{A'}, X_v, X_{A'}, X_v, X_C \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \mid X_C \\
\xRightarrow{\text{Left Decomposition, } A \subseteq A' \dot\cup \{v\} \dot\cup C} \quad & X_A \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \mid X_C.
\end{aligned}
$$

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:

$$
\begin{aligned}
A \overset{d}{\underset{G}{\perp}} J \cup B \mid C \quad & \xRightarrow{\text{Left Decomposition}} \quad A' \overset{d}{\underset{G}{\perp}} J \cup B \mid C \\
& \xRightarrow{\text{Right Decomposition}} \quad A' \overset{d}{\underset{G}{\perp}} J \cup B' \mid C \\
& \xRightarrow{\text{marginalization, } v \notin A' \cup J \cup B' \cup C} \quad A' \overset{d}{\underset{G^{\setminus \{v\}}}{\perp}} J \cup B' \mid C \\
& \xRightarrow{\text{induction (IND)}} \quad X_{A'} \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{B'} \mid X_C. \qquad (\#1')
\end{aligned}
$$

Again with $D = \mathrm{Pa}^G(v)$ we get:

$$
\begin{aligned}
A \overset{d}{\underset{G}{\perp}} J \cup B \mid C \quad & \xRightarrow{\text{Left Decomposition}} \quad A' \overset{d}{\underset{G}{\perp}} J \cup B \mid C \\
& \xRightarrow{\text{Right Decomposition}} \quad A' \overset{d}{\underset{G}{\perp}} J \cup B' \cup \{v\} \mid C \\
& \xRightarrow{\text{Right Weak Union}} \quad A' \overset{d}{\underset{G}{\perp}} J \cup \{v\} \mid B' \dot\cup C \\
& \xRightarrow{(\bullet), \text{ see below}} \quad A' \overset{d}{\underset{G}{\perp}} J \cup D \mid B' \dot\cup C \\
& \xRightarrow{\text{induction (IND)}} \quad X_{A'} \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_D \mid X_{B' \dot\cup C} \\
& \xRightarrow{B' \dot\cup C} \quad X_{A'} \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_D \mid X_{B'}, X_C. \qquad (\#2')
\end{aligned}
$$

$(\bullet)$ holds since every $(B' \dot\cup C)$-open walk $\cdots \ast\!\!\rightarrowtail\!\!\ast\, w$ 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 $\cdots \ast\!\!\rightarrowtail\!\!\ast\, w \rightarrowtail v$ 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:

$$X_v \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{\mathrm{Pred}^G_<(v)} \mid X_D.$$

With this and $A' \cup J \cup B' \cup C \subseteq \mathrm{Pred}^G_<(v)$ we get the implications:

$$\begin{aligned}
& X_v \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{\mathrm{Pred}^G_<(v)} \mid X_D \\
\xRightarrow{\text{Right Decomposition}} \quad & X_v \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{A'}, X_{B'}, X_C \mid X_D \\
\xRightarrow{\text{Right Weak Union}} \quad & X_v \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{A'} \mid X_{B'}, X_C, X_D \\
\xRightarrow{\text{Flipped Left Cross Contraction, (\#2')}} \quad & X_{A'} \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_D, X_v \mid X_{B'}, X_C \\
\xRightarrow{\text{Right Decomposition}} \quad & X_{A'} \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_v \mid X_{B'}, X_C \\
\xRightarrow{\text{Right Contraction, (\#1')}} \quad & X_{A'} \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_{B'}, X_v \mid X_C \\
\xRightarrow{X_J\text{-Inverted Right Decomposition}} \quad & X_{A'} \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_J, X_{B'}, X_v, X_C \mid X_C \\
\xRightarrow{\text{Right Decomposition, } B \subseteq B' \dot\cup \{v\} \dot\cup C} \quad & X_{A'} \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \mid X_C. \qquad (\#3')
\end{aligned}$$

By Redundancy we have:

$$X_{A'}, X_C \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \mid X_{A'}, X_C.$$

With this we get the implications:

$$\begin{aligned}
& X_{A'}, X_C \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \mid X_{A'}, X_C \\
\xRightarrow{\text{Left Contraction, (\#3')}} \quad & X_{A'}, X_{A'}, X_C \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \mid X_C. \\
\xRightarrow{\text{Left Decomposition, } A \subseteq A' \dot\cup C} \quad & X_A \underset{P(X_V \mid \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \mid X_C.
\end{aligned}$$

This shows the claim in case 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.

$$A \overset{d}{\underset{G}{\perp}} J \cup B \,|\, C \xRightarrow{\text{Left Decomposition}} A' \overset{d}{\underset{G}{\perp}} J \cup B \,|\, C$$

$$\xRightarrow{\text{Right Decomposition}} A' \overset{d}{\underset{G}{\perp}} J \cup B' \,|\, C$$

$$\xRightarrow{C=C' \dot\cup \{v\}} A' \overset{d}{\underset{G}{\perp}} J \cup B' \,|\, C' \dot\cup \{v\}$$

We now claim that:

$$A' \overset{d}{\underset{G}{\perp}} J \cup B' \,|\, C' \dot\cup \{v\}$$

implies that one of the following statements holds:

$$A' \dot\cup \{v\} \overset{d}{\underset{G}{\perp}} J \cup B' \,|\, C' \qquad \vee \qquad A' \overset{d}{\underset{G}{\perp}} J \cup (B' \dot\cup \{v\}) \,|\, C'.$$

Assume the contrary:

$$A' \dot\cup \{v\} \overset{d}{\underset{G}{\not\perp}} J \cup B' \,|\, C' \qquad \wedge \qquad A' \overset{d}{\underset{G}{\not\perp}} J \cup (B' \dot\cup \{v\}) \,|\, C'.$$

So there exist shortest C'-open walks π_1 and π_2 in G:

$$\pi_1 : \quad A' \cup \{v\} \ni u_0 \ast\!\!\rightarrow \cdots \ast\!\!\rightarrow u_k \in J \cup B',$$

and:

$$\pi_2 : \quad A' \ni w_0 \ast\!\!\rightarrow \cdots \ast\!\!\rightarrow w_m \in J \cup (B' \dot\cup \{v\}).$$

So all colliders of π_1 and π_2 are in C' and all non-colliders 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:

$$A' \overset{d}{\underset{G}{\perp}} J \cup B' \,|\, C' \dot\cup \{v\}.$$

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:

$$\pi_1 : \quad v \leftarrowtail u_1 \ast\!\!\rightarrow \cdots \ast\!\!\rightarrow u_k,$$

and:

$$\pi_2 : \quad w_0 \ast\!\!\rightarrow \cdots \ast\!\!\rightarrow w_{m-1} \rightarrowtail v,$$

with $u_1, w_{m-1} \in D = \mathrm{Pa}^G(v)$. Then the following walk:

$$A' \ni w_0 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast w_{m-1} \rightarrowtail v \leftarrowtail u_1 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast u_k \in J \cup B',$$

is a $(C' \dot\cup \{v\})$-open walk from A' to $J \cup B'$, in contradiction to:

$$A' \overset{d}{\underset{G}{\perp}} J \cup B' \,|\, C' \dot\cup \{v\}.$$

So the claim:

$$A' \dot\cup \{v\} \overset{d}{\underset{G}{\perp}} J \cup B' \,|\, C' \qquad \vee \qquad A' \overset{d}{\underset{G}{\perp}} J \cup (B' \dot\cup \{v\}) \,|\, C',$$

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 with input variables)**.** *Consider a causal Bayesian network with input variables (I-CBN) given by* $(G, P(X_V|\,\mathrm{do}(X_J)))$ *with CDAG:* $G = (J, V, E)$ *and Markov kernels:* $P_v\left(X_v|\,\mathrm{do}\left(X_{\mathrm{Pa}^G(v)}\right)\right)$ *for* $v \in V$*. Now let* $W \subseteq J \cup V$ *any subset. Then we define the* intervened causal Bayesian network with input variables *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:* $P_v\left(X_v|\,\mathrm{do}\left(X_{\mathrm{Pa}^G(v)}\right)\right)$ *for* $v \in V \setminus W$*.*

Its observational Markov kernel is then:

$$P\left(X_{V\setminus W}|\,\mathrm{do}\left(X_{J\cup W}\right)\right) = \bigotimes_{v\in V\setminus W} P_v\left(X_v|\,\mathrm{do}\left(X_{\mathrm{Pa}^G(v)}\right)\right).$$

Note that if $v \in V \setminus W$ *then* $\mathrm{Pa}^G(v) = \mathrm{Pa}^{G_{\mathrm{do}(W)}}(v)$*.*

Remark 4.3.2 (Hard intervention on causal Bayesian network with input variables and with latent variables)**.** *We define the interventions on IL-CBN the same way as above, 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. Soft Interventions on Causal Bayesian Networks

Remark 4.3.1 (Modelling soft interventions on causal Bayesian network with input variables)**.** *Consider a causal Bayesian network with input variables given by* $(G, P(X_V|\operatorname{do}(X_J)))$ *with CDAG:* $G = (J, V, E)$ *and Markov kernels:* $P_v\left(X_v|\operatorname{do}\left(X_{\mathrm{Pa}^G(v)}\right)\right)$ *for* $v \in V$*.*
Let $W \subseteq J \cup V$*. If we want to model a soft intervention on variables* X_w *for* $w \in W \setminus J$ *then we would introduce soft intervention nodes* $I_w \rightarrowtail w$ *for* $w \in W \setminus J$*, which come with new input variables* X_{I_w}*, and replace the Markov kernel:*

$$P_w\left(X_w|\operatorname{do}\left(X_{\mathrm{Pa}^G(w)}\right)\right)$$

for $w \in W \setminus J$ *by one that models the dependence on the soft intervention variables properly:*

$$P_w\left(X_w|\operatorname{do}\left(X_{\mathrm{Pa}^G(w)}, X_{I_w}\right)\right).$$

So the soft interventional causal Bayesian network with input variables *w.r.t.* W *would then have:*

1. *CDAG:* $G_{\operatorname{do}(I_W)} = (J \dot\cup \{I_w\,|w \in W\}, V, E \dot\cup \{I_w \rightarrowtail w\,|w \in W \setminus J\})$*, and:*
2. *Markov kernels:*

   $P_v\left(X_v|\operatorname{do}\left(X_{\mathrm{Pa}^G(v)}\right)\right)$ *for* $v \in V \setminus W$*, and:*

   $P_w\left(X_w|\operatorname{do}\left(X_{\mathrm{Pa}^G(w)}, X_{I_w}\right)\right)$ *for* $w \in W \setminus J$*.*

Note that $\mathrm{Pa}^{G_{\operatorname{do}(I_W)}}(w) = \mathrm{Pa}^G(w) \dot\cup \{I_w\}$ *for* $w \in W \setminus J$ *and* $\mathrm{Pa}^{G_{\operatorname{do}(I_W)}}(v) = \mathrm{Pa}^G(v)$ *for* $v \in V \setminus W$*.*

Remark 4.3.2 (Modelling hard interventions as soft interventions)**.** *It is sometimes beneficial to model hard interventions as soft interventions. Let the setting be like in 4.3.1. When we model hard interventions as soft interventions 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. $X_{I_w} := \mathrm{pr}_{\mathcal{X}_{I_w}}$,
3. $P_w\left(X_w \in A|\operatorname{do}\left(X_{\mathrm{Pa}^G(w)} = x_{\mathrm{Pa}^G(w)}, X_{I_w} = x_{I_w}\right)\right) :=$

$$\begin{cases} P_w\left(X_w \in A|\operatorname{do}\left(X_{\mathrm{Pa}^G(w)} = x_{\mathrm{Pa}^G(w)}\right)\right), & \text{if} \quad x_{I_w} = \star, \\ \delta(X_w \in A|X_w = x_{I_w}) = \mathbb{1}_A(x_{I_w}), & \text{if} \quad x_{I_w} \neq \star. \end{cases}$$

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:* $\operatorname{do}(X_w = x_{I_w})$*.*

Remark 4.3.3. *Again, we can do all the above also with causal Bayesian network with input variables and with latent variables, but allow only* W *with* $W \cap U = \emptyset$*.*

4.3.3. Marginalization of Causal Bayesian Networks

Definition 4.3.1 (Marginalization of causal Bayesian network with input variables and with latent variables)**.** *Consider a causal Bayesian network with input variables and with latent variables (IL-CBN):*

$$\left(G^+ = (J, V, U, E^+),\ P\left(X_{V \dot\cup U} | \operatorname{do}(X_J)\right)\right).$$

Let $W \subseteq V$ be a subset. We then define the marginalized IL-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 observational Markov kernel marginalizes to:

$$P\left(X_{V \setminus W} | \operatorname{do}(X_J)\right),$$

and the observational CADMG becomes:

$$(G^+)^{\setminus (U \dot\cup W)} = G^{\setminus W},$$

i.e. the marginalized G w.r.t. W.

4.4. Representations of Causal Bayesian Networks

4.4.1. Interventional Equivalence

Definition 4.4.1. *Consider two causal Bayesian network with input variables and with latent variables (IL-CBNs):*

$$M_1 = \left(G_1^+ = (J_1, V_1, U_1, E_1^+),\ P_1\left(X_{V_1}, X_{U_1} | \operatorname{do}(X_{J_1})\right)\right),$$

$$M_2 = \left(G_2^+ = (J_2, V_2, U_2, E_2^+),\ P_2\left(X_{V_2}, X_{U_2} | \operatorname{do}(X_{J_2})\right)\right).$$

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* $W \subseteq J \cup V$ *we have the equality of the Markov kernels:*

$$P_1\left(X_{V \setminus W} | \operatorname{do}(X_{J \cup W})\right) = P_2\left(X_{V \setminus W} | \operatorname{do}(X_{J \cup W})\right).$$

4.4.2. Structural Causal Model Representation

Theorem 4.4.1 (Structural causal model representation of IL-CBNs)**.** *Consider a causal Bayesian network with input variables and with latent variables (IL-CBN):*

$$M = \left(G^+ = (J, V, U, E^+),\ P\left(X_V, X_U | \operatorname{do}(X_J)\right)\right).$$

Then there exists an interventionally equivalent IL-CBN:

$$\tilde{M} = \left(\tilde{G}^+ = (J, V, \tilde{U}, \tilde{E}^+),\ \tilde{P}\left(X_V, X_{\tilde{U}} | \operatorname{do}(X_J)\right)\right),$$

such that:

1. $\mathrm{Pa}^{\tilde{G}^+}(\tilde{U}) = \emptyset$,
2. $\mathrm{Ch}^{\tilde{G}^+}(\tilde{U}) = V$,
3. $\mathrm{Ch}^{\tilde{G}^+}(u_1) \not\subseteq \mathrm{Ch}^{\tilde{G}^+}(u_2)$ *for* $u_1, u_2 \in \tilde{U}$ *with* $u_1 \neq u_2$,
4. *all Markov kernels* P_v *for* $v \in V$ *are deterministic, i.e. of form* $\delta(R_v | X_{\mathrm{Pa}^{\tilde{G}^+}(v)})$ *for some deterministic function* R_v*, i.e.:* $X_v = R_v(X_{\mathrm{Pa}^{\tilde{G}^+}(v)})$,
5. $G := (G^+)^{\backslash U} = (\tilde{G}^+)^{\backslash \tilde{U}}$.

Furthermore, if we are willing to use measurable embeddings/isomorphisms: $\mathcal{X}_u \hookrightarrow [0,1]$*, then we can further restrict for* $u \in \tilde{U}$ *to:*

6. $\mathcal{X}_u \cong [0,1]$
7. $P_u(X_u)$ *is the uniform distribution on* $[0,1]$.

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 theorem 2.7.4:

$$P_v(X_v | \operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)})) = \delta(R_v | X_{\bar{v}}, X_{\mathrm{Pa}^{G^+}(v)}) \circ \bar{P}_{\bar{v}}(X_{\bar{v}}).$$

We now put:

$$\bar{U} := U \dot{\cup} \{\bar{v} \,|\, v \in V \cup U\}, \qquad \bar{E}^+ := E^+ \dot{\cup} \{\bar{v} \rightarrowtail v \,|\, v \in V \cup U\},$$

and to get $\bar{M}$ we add the $\bar{P}_{\bar{v}}$ to M and replace P_v for $v \in V \cup U$ by the deterministic one given by:

$$\bar{P}_v(X_v \in A | \operatorname{do}(X_{\bar{v}}, X_{\mathrm{Pa}^{G^+}(v)})) := \delta(R_v \in A | X_{\bar{v}}, X_{\mathrm{Pa}^{G^+}(v)}).$$

Then $\bar{G}^+$ clearly marginalizes to G^+ (when we marginalize out all the $\bar{v}$ again) and the marginal of:

$$\bar{P}_v(X_v \in A | \operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)}), X_{\bar{v}}) \otimes \bar{P}_{\bar{v}}(X_{\bar{v}}),$$

in the defining product of the joint Markov kernel is $P_v(X_v|\operatorname{do}(X_{\operatorname{Pa}^{G^+}(v)}))$ 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 $u \in U$:

$$\begin{aligned}&\int_{\mathcal{X}_u} \bigotimes_{v \in \operatorname{Ch}^{\bar{G}^+}(u)} \bar{P}_v(X_v|\operatorname{do}(X_{\operatorname{Pa}^{\bar{G}^+}(v)\setminus\{u\}}, X_u = x_u))\, \delta(R_u \in dx_u | X_{\operatorname{Pa}^{\bar{G}^+}(u)})\\ &= \bigotimes_{v \in \operatorname{Ch}^{\bar{G}^+}(u)} \bar{P}_v(X_v|\operatorname{do}(X_{\operatorname{Pa}^{\bar{G}^+}(v)\setminus\{u\}}, X_u = R_u(X_{\operatorname{Pa}^{\bar{G}^+}(u)}))),\end{aligned}$$

which is again a product (only) because we marginalized a deterministic Markov kernel out. So we define:

$$\hat{P}_v(X_v|\operatorname{do}(X_{\operatorname{Pa}^{\hat{G}^+}(v)})) := \bar{P}_v(X_v|\operatorname{do}(X_{\operatorname{Pa}^{\bar{G}^+}(v)\setminus\{u\}}, X_u = R_u(X_{\operatorname{Pa}^{\bar{G}^+}(u)}))),$$

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:

$$\operatorname{Pa}^{\hat{G}^+}(v) := \operatorname{Pa}^{\bar{G}^+}(v) \setminus \{u\} \cup \operatorname{Pa}^{\bar{G}^+}(u),$$

i.e. the CDAG from $(\bar{G}^+)^{\setminus U}$ where we removed all bi-directed edges, $\hat{U} = \bar{U} \setminus U$. Then note that for $u \in U$ we have:

$$\operatorname{Ch}^{\bar{G}^+}(u) = \operatorname{Ch}^{(\bar{G}^+)^{\setminus U}}(\bar{u}).$$

This implies that we recover the removed bi-directed edges if we further marginalize out all the $\bar{u}$, i.e.:

$$(\hat{G}^+)^{\setminus \hat{U}} = (\bar{G}^+)^{\setminus \bar{U}} = (G^+)^{\setminus U} = G.$$

With this we already achieved all points 1.-7. from the statement except point 3.
Step 3. First we can marginalize out all nodes $u \in \hat{U}$ with $\operatorname{Ch}^{\hat{G}^+}(u) = \emptyset$. For those u we have:

$$\hat{P}(X_V, X_{\hat{U}}|\operatorname{do}(X_J)) = \hat{P}_u(X_u|\operatorname{do}(X_{\operatorname{Pa}^{\hat{G}^+}(u)})) \otimes \hat{P}(X_V, X_{\hat{U}\setminus\{u\}}|\operatorname{do}(X_J)).$$

So marginalizing out X_u does not interfere with the rest of the Markov kernels.
Step 4. If then $\operatorname{Ch}^{\tilde{G}^+}(u_1) \subseteq \operatorname{Ch}^{\tilde{G}^+}(u_2)$ for some $u_1, u_2 \in \hat{U}$ then consider the space $\mathcal{X}_{\tilde{u}_{1,2}} := \mathcal{X}_{u_1} \times \mathcal{X}_{u_2}$, the variable $X_{\tilde{u}_{1,2}} := (X_{u_1}, X_{u_2})$. Then every Markov kernel dependent on X_{u_1} or X_{u_2} can be written as a Markov kernel dependent on $X_{\tilde{u}_{1,2}}$. In the graph we replace both u_1 and u_2 by a single node $\tilde{u}_{1,2}$ with the same edges as u_2. Repeating this procedure then gives the mutual non-inclusivity. It is clear that: $(\tilde{G}^+)^{\setminus \tilde{U}} = (\hat{G}^+)^{\setminus \hat{U}}$ as we remove all elements from $\tilde{U}$, $\hat{U}$, resp., and introduce in both cases bi-directed edges between all children of u_2.
Step 5. We can use the same arguments as in step 1 and 2 to make $\mathcal{X}_{\tilde{u}} \cong [0,1]$ and $P_{\tilde{u}}(X_{\tilde{u}})$ the uniform distribution (while just composing deterministic Markov kernels).
This shows all the claims. Finally one can convince oneself that at each step we get an interventionally equivalent IL-CBN to the step before. □

4.4.3. Standard Form of Causal Bayesian Networks

Definition/Theorem 4.4.1 (Standard forms of IL-CBNs)**.** *Consider a causal Bayesian network with input variables and with latent variables (IL-CBN):*

$$M = \left(G^+ = (J, V, U, E^+),\ P\left(X_V, X_U | \operatorname{do}\left(X_J\right)\right)\right).$$

Then there exists an interventionally equivalent IL-CBN:

$$\tilde{M} = \left(\tilde{G}^+ = (J, V, \tilde{U}, \tilde{E}^+),\ \tilde{P}\left(X_V, X_{\tilde{U}} | \operatorname{do}\left(X_J\right)\right)\right),$$

such that for every $u, u_1, u_2 \in \tilde{U}$ with $u_1 \neq u_2$:

1. $\mathrm{Pa}^{\tilde{G}^+}(u) = \emptyset$,
2. $\#\mathrm{Ch}^{\tilde{G}^+}(u) \geq 2$,
3. $\mathrm{Ch}^{\tilde{G}^+}(u_1) \not\subseteq \mathrm{Ch}^{\tilde{G}^+}(u_2)$,
4. $G := (G^+)^{\backslash U} = (\tilde{G}^+)^{\backslash \tilde{U}}$.

Furthermore, the CDAG $\tilde{G}^+$ is uniquely determined by G^+. We will call such a IL-CBN a standard form *of the original IL-CBN.*
Furthermore, if we are willing to use measurable embeddings/isomorphisms: $\mathcal{X}_u \hookrightarrow [0,1]$, then we can further restrict for $u \in \tilde{U}$ to:

5. $\mathcal{X}_u \cong [0,1]$
6. $\tilde{P}_u(X_u)$ *is the uniform distribution on* $[0,1]$.

Note that in any case the Markov kernels dependent on X_U might not be unique as we can always transform $[0,1]$ to $[0,1]$ in strange ways.

Proof. This follows from the SCM representation of LCCMNs, theorem 4.4.1, by marginalizing over 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:

$$P_v(X_v | \operatorname{do}(X_{\mathrm{Pa}^{\tilde{G}^+}(v)\backslash\{u\}}, X_u)) \otimes P_u(X_u),$$

by the composition:

$$P_v(X_v | \operatorname{do}(X_{\mathrm{Pa}^{\tilde{G}^+}(v)\backslash\{u\}}, X_u)) \circ P_u(X_u),$$

which might not be deterministic anymore. □

Remark 4.4.2 (Marginalizations and hard interventions on standard forms)**.** *Let the following IL-CBN be in standard form:*

$$\left(G^+ = (J, V, U, E^+),\ P\left(X_{V \cup U} | \operatorname{do}\left(X_J\right)\right)\right).$$

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 standard form of that procedure.
Similarly we could post-process hard interventions with standardization steps.

Causal Bayesian Networks - Continued

4.4.1. Do-Calculus

Remark 4.4.1 (Recap)**.** *Consider an IL-BCN:*

$$M = \left(G^+ = (J, V, U, E^+), \quad \left(P_v\left(X_v | \operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)})\right)\right)_{v \in V \cup U}\right).$$

Then we get the joint Markov kernel over all input, observed and unobserved output variables as follows:

$$P\left(X_V, X_U, X_J | \operatorname{do}(X_J)\right) := \bigotimes_{v \in U \cup V} P_v\left(X_v | \operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)})\right) \otimes \bigotimes_{j \in J} \delta(X_j | X_j).$$

Further, for $D \subseteq J \cup V$ and $C \subseteq V \setminus D$ we get the combined soft and hard interventions:

$$P\left(X_{V \setminus D}, X_U, X_{J \cup D}, X_{I_C} | \operatorname{do}(X_{I_C}, X_{J \cup D})\right) :=$$

$$\bigotimes_{v \in V \setminus (C \cup D)} P_v\left(X_v | \operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)})\right) \otimes \bigotimes_{v \in C} P_v\left(X_v | \operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)}, X_{I_v})\right) \otimes$$

$$\bigotimes_{v \in U} P_v\left(X_v | \operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)})\right) \otimes \bigotimes_{j \in J \cup D} \delta(X_j | X_j) \otimes \bigotimes_{v \in C} \delta(X_{I_v} | X_{I_v}),$$

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, $v \in C$:

$$P_v\left(X_v | \operatorname{do}\left(X_{\mathrm{Pa}^{G^+}(v)}, X_{I_v} = x_{I_v}\right)\right) := \begin{cases} P_v\left(X_v | \operatorname{do}\left(X_{\mathrm{Pa}^{G^+}(v)}\right)\right), & \text{if} \quad x_{I_v} = \star, \\ \delta(X_v | X_v = x_{I_v}), & \text{if} \quad x_{I_v} \neq \star. \end{cases}$$

Finally we can also marginalize (i.e. integrating out) and condition to get:

$$P\left(X_A | X_B, \operatorname{do}(X_{J \cup D}, X_{I_C})\right),$$

for any $A, B, C, D \subseteq J \cup V$.
For more suggestive formulas later on we also freely permute the order of symbols behind the conditioning line, e.g.:

$$P\left(X_A | \operatorname{do}(X_F), X_B, \operatorname{do}(X_D)\right) := P\left(X_A | X_B, \operatorname{do}(X_D, X_F)\right).$$

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 do the intervention (*do*) first and afterwards condition on the further variables.*
We will also make use of the following CADMG:

$$G_{\operatorname{do}(I_C, D)} = (G^+_{\operatorname{do}(I_C, D)})^{\setminus U}.$$

W.l.o.g. we can assume: $J \subseteq D$ and $C \cap D = \emptyset$.

Corollary 4.4.2 (Do-calculus)**.** *Consider an IL-BCN:*

$$M = \left(G^+ = (J, V, U, E^+), \quad \left(P_v\left(X_v | \operatorname{do}(X_{\operatorname{Pa}^{G^+}(v)})\right)\right)_{v \in V \cup U}\right).$$

Let $A, B, C \subseteq V$ and $D \subseteq J \cup V$ be 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:*

$$A \underset{G_{\operatorname{do}(D)}}{\overset{d}{\perp}} B \,|\, C \cup D,$$

then there exists a Markov kernel:

$$P\left(X_A | \cancel{X_B}, X_C, \operatorname{do}(X_{D \cup J})\right) : \mathcal{X}_C \times \mathcal{X}_D \dashrightarrow \mathcal{X}_A$$

that is a version of:

$$P\left(X_A | X_{B_2}, X_C, \operatorname{do}(X_{D \cup J})\right),$$

for every subset $B_2 \subseteq B$ simultaneously. In short we could write:

$$P\left(X_A | X_B, X_C, \operatorname{do}(X_{D \cup J})\right) = P\left(X_A | X_C, \operatorname{do}(X_D)\right).$$

Note that this Markov kernel is only dependent on x_C and x_D, but constant in $x_{J \setminus D}$. Such a Markov kernel is unique up to $P\left(X_B, X_C | \operatorname{do}(X_{D \cup J})\right)$-null set.

2. *Action/observation exchange: If we have:*

$$A \underset{G_{\operatorname{do}(I_B, D)}}{\overset{d}{\perp}} I_B \,|\, B \cup C \cup D,$$

then there exists a Markov kernel:

$$P\left(X_A | \cancel{\operatorname{do}}(X_B), X_C, \operatorname{do}(X_{D \cup J})\right) : \mathcal{X}_B \times \mathcal{X}_C \times \mathcal{X}_D \dashrightarrow \mathcal{X}_A$$

that is a version of:

$$P\left(X_A | X_{B_1}, \operatorname{do}(X_{B_2}), X_C, \operatorname{do}(X_{D \cup J})\right),$$

for every decomposition: $B = B_1 \dot{\cup} B_2$, simultaneously. In short we could write:

$$P\left(X_A | \operatorname{do}(X_B), X_C, \operatorname{do}(X_{D \cup J})\right) = P\left(X_A | X_B, X_C, \operatorname{do}(X_{D \cup J})\right),$$

or even

$$P\left(X_A | \operatorname{do}(X_B), X_C, \operatorname{do}(X_D)\right) = P\left(X_A | X_B, X_C, \operatorname{do}(X_D)\right)$$

since these Markov kernels are constant in $x_{J \setminus D}$. Such a Markov kernel is unique up to $P\left(X_B, X_C | \operatorname{do}(X_{I_B}, X_{D \cup J})\right)$-null set.

3. *Insertion/deletion of action:* *If we have:*

$$A \underset{G_{\mathrm{do}(I_B,D)}}{\overset{d}{\perp}} I_B \,|\, C \cup D,$$

then there exists a Markov kernel:

$$P\left(X_A|\cancel{\mathrm{do}(X_B)}, X_C, \mathrm{do}(X_{D\cup J})\right) : \mathcal{X}_C \times \mathcal{X}_D \dashrightarrow \mathcal{X}_A$$

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. In short we could write:

$$P\left(X_A|\,\mathrm{do}(X_B), X_C, \mathrm{do}(X_{D\cup J})\right) = P\left(X_A|X_C, \mathrm{do}(X_D)\right).$$

Note that this Markov kernel is only dependent on x_C and x_D, but constant in x_J. Such a Markov kernel is unique up to $P\left(X_C|\,\mathrm{do}(X_{I_B}, X_{D\cup J})\right)$-null set.

Proof. We make use of the global Markov property (GMP), theorem 4.2.1.

Point 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:

$$P\left(X_A, X_B, X_C|\,\mathrm{do}(X_{D\cup J})\right) = Q(X_A|X_C, X_D) \otimes P\left(X_B, X_C|\,\mathrm{do}(X_{D\cup J})\right),$$

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:

$$P\left(X_A|X_B, X_C, \mathrm{do}(X_{D\cup J})\right).$$

If we marginalize out X_{B_1} for any decomposition $B = B_1 \dot\cup B_2$ in the above factorization we also get:

$$P\left(X_A, X_C, X_{B_2}|\,\mathrm{do}(X_{D\cup J})\right) = Q(X_A|X_C, X_D) \otimes P\left(X_C, X_{B_2}|\,\mathrm{do}(X_{D\cup J})\right),$$

showing that $Q(X_A|X_C, X_D)$ is also a version of:

$$P\left(X_A|X_{B_2}, X_C, \mathrm{do}(X_{D\cup J})\right).$$

In particular, this holds for $B_2 = \emptyset$.

Point 2.) The assumption:

$$A \underset{G_{\mathrm{do}(I_B, D)}}{\overset{d}{\perp}} I_B \,|\, B \cup C \cup D,$$

implies the conditional independence by GMP 4.2.1:

$$X_A \underset{P(X_V \,|\, \mathrm{do}(X_{I_B}, X_{D \cup J}))}{\perp\!\!\!\perp} X_{I_B} \,|\, X_B, X_C, X_D.$$

So we have the following factorization:

$$P\left(X_A, X_B, X_C | \,\mathrm{do}(X_{I_B}, X_{D\cup J})\right) = Q(X_A | X_B, X_C, X_D) \otimes P\left(X_B, X_C | \,\mathrm{do}(X_{I_B}, X_{D\cup J})\right),$$

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 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:

$$\begin{aligned}
&P\left(X_A, X_B, X_C | \,\mathrm{do}(X_{B_2} = x_{B_2}, X_{D\cup J})\right) \\
&= P\left(X_A, X_B, X_C | \,\mathrm{do}(X_{I_{B_1}} = \star, X_{I_{B_2}} = x_{B_2}, X_{D\cup J})\right) \\
&= Q(X_A | X_B, X_C, X_D) \otimes P\left(X_B, X_C | \,\mathrm{do}(X_{I_{B_1}} = \star, X_{I_{B_2}} = x_{B_2}, X_{D\cup J})\right), \\
&= Q(X_A | X_B, X_C, X_D) \otimes P\left(X_B, X_C | \,\mathrm{do}(X_{B_2} = x_{B_2}, X_{D\cup J})\right).
\end{aligned}$$

So we get:

$$P\left(X_A, X_B, X_C | \,\mathrm{do}(X_{B_2}, X_{D\cup J})\right) = Q(X_A | X_B, X_C, X_D) \otimes P\left(X_B, X_C | \,\mathrm{do}(X_{B_2}, X_{D\cup J})\right),$$

where we can marginalize out the deterministic X_{B_2}:

$$P\left(X_A, X_{B_1}, X_C | \,\mathrm{do}(X_{B_2}, X_{D\cup J})\right) = Q(X_A | X_B, X_C, X_D) \otimes P\left(X_{B_1}, X_C | \,\mathrm{do}(X_{B_2}, X_{D\cup J})\right).$$

This implies that $Q(X_A|X_B, X_C, X_D)$ is a version of the conditional Markov kernel:

$$P\left(X_A | X_{B_1}, X_C, \mathrm{do}(X_{B_2}, X_{D\cup J})\right),$$

for every decomposition: $B = B_1 \dot\cup B_2$, simultaneously, in particular for the two cases $B_2 = \emptyset$ and $B_2 = B$.

Point 3.) The assumption:

$$A \underset{G_{\mathrm{do}(I_B, D)}}{\overset{d}{\perp}} I_B \,|\, C \cup D,$$

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:

$$P\left(X_A, X_C|\,\mathrm{do}(X_{I_B}, X_{D\cup J})\right) = Q(X_A|X_C, X_D) \otimes P\left(X_C|\,\mathrm{do}(X_{I_B}, X_{D\cup J})\right),$$

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:

$$P\left(X_A, X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J})\right) = Q(X_A|X_C, X_D) \otimes P\left(X_C|\,\mathrm{do}(X_{B_2}, X_{D\cup J})\right),$$

which shows that $Q(X_A|X_C, X_D)$ is a version of the conditional Markov kernel:

$$P\left(X_A|X_C, \mathrm{do}(X_{B_2}, X_{D\cup J})\right)$$

for every decomposition: $B = B_1 \dot\cup B_2$, simultaneously, in particular for the two corner cases: $B_2 = \emptyset$ and $B_2 = B$.

□

4.4.2. Identifying Causal Effects

Motivation 4.4.1. *Consider an IL-BCN:*

$$M = \left(G^+ = (J, V, U, E^+), \quad \left(P_v\left(X_v|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)})\right)\right)_{v\in V\cup U}\right).$$

*For simplicity assume that there are no input variables, i.e. $J = \emptyset$. Then the joint distribution is "*do*-free" and given as:*

$$P(X_V, X_U) = \prod_{v\in U\cup V} P_v\left(X_v|\,\mathrm{do}(X_{\mathrm{Pa}^{G^+}(v)})\right),$$

with observational distribution as its marginal: $P(X_V)$.
We also have all the interventional distributions for $W \subseteq V$*:*

$$P(X_{V\setminus W}, X_U | \operatorname{do}(X_W)) = \prod_{v \in U \cup V \setminus W} P_v \left(X_v | \operatorname{do}(X_{\operatorname{Pa}^{G^+}(v)}) \right),$$

with marginals: $P(X_{V\setminus W} | \operatorname{do}(X_W))$.
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:* $P(X_{V\setminus W} | \operatorname{do}(X_W))$ *from data, e.g. if vaccination makes people immune to a disease, we typically would need to perform an interventional study where we intervene on the variables* X_W *on different values. This usually requires expensive, time-consuming randomized control trials with an own comparison group for each possible values 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* $\operatorname{do}(W)$ *to expressions only involving* $\operatorname{do}(D)$ *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: $v_1 \rightarrowtail v_2$ *we have that:*

$$P(X_2 | \operatorname{do}(X_1)) = P(X_2 | X_1),$$

which can be estimated using observational data only, e.g. via supervised learning.
So the question is now: Assuming that the causal graph is known, under which circumstances is a causal effect $P(X_A | \operatorname{do}(X_B))$ *already determined by the observational distribution* $P(X_V)$*? When can causal effects be identified via distributions that have less interventions in them?*

Notation 4.4.2. *Consider an IL-BCN:*

$$M = \left(G^+ = (J, V, U, E^+), \quad \left(P_v \left(X_v | \operatorname{do}(X_{\operatorname{Pa}^{G^+}(v)}) \right) \right)_{v \in V \cup U} \right).$$

We are interested in estimating the conditional causal effect:

$$P(X_A | X_C, \operatorname{do}(X_B, X_D)),$$

but we only have data from:

$$P(X_V | X_C, \operatorname{do}(X_D)).$$

The following index sets will have the following roles:

1. A*: the outcome variables of interest.*
2. B*: the treatment or intervention variables.*
3. C*: general conditional (context) variables under which the data was collected.*

4. *D: general interventional (context) variables that were set by the experimenter, $J \subseteq D$.*
5. *F_0: core adjustment variables, i.e. features that were measured.*
6. *F_1: additional measured adjustment variables.*
7. *$F = F_0 \cup F_1$.*
8. *H: additional unobserved variables.*

Theorem 4.4.3 (General adjustment formula)**.** *Consider an IL-BCN:*

$$M = \left(G^+ = (J, V, U, E^+), \quad \left(P_v\left(X_v | \operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)})\right)\right)_{v \in V \cup U}\right).$$

Assume that all the following conditions hold in the graphs $G^+_{\operatorname{do}(I_B, D)}$:

$$(F_0 \cup H) \underset{G^+_{\operatorname{do}(I_B,D)}}{\overset{d}{\perp}} I_B \,|(C \cup D), \tag{19}$$

$$A \underset{G^+_{\operatorname{do}(I_B,D)}}{\overset{d}{\perp}} (F_1 \cup I_B)\,|(B \cup F_0 \cup H \cup C \cup D), \tag{20}$$

$$H \underset{G^+_{\operatorname{do}(I_B,D)}}{\overset{d}{\perp}} B\,|(F \cup C \cup I_B \cup D). \tag{21}$$

Then we have the adjustment formula:

$$P(X_A|X_C, \operatorname{do}(X_B, X_D)) = P(X_A|X_B, X_C, X_F, \operatorname{do}(X_D)) \circ P(X_F|X_C, \operatorname{do}(X_D)) \qquad a.s.$$

Proof.

$$P(X_A|X_C, \operatorname{do}(X_B, X_D)) \tag{22}$$
$$= P(X_A|X_{F_0}, X_H, X_C, \operatorname{do}(X_B, X_D)) \circ P(X_{F_0}, X_H|X_C, \operatorname{do}(X_B, X_D)) \tag{23}$$
$$\overset{20}{=} P(X_A|X_{F_0}, X_H, X_C, X_B, \operatorname{do}(X_D)) \circ P(X_{F_0}, X_H|X_C, \operatorname{do}(X_B, X_D)) \tag{24}$$
$$\overset{19}{=} P(X_A|X_{F_0}, X_H, X_C, X_B, \operatorname{do}(X_D)) \circ P(X_{F_0}, X_H|X_C, \operatorname{do}(X_D)) \tag{25}$$
$$= P(X_A|X_{F_0}, X_H, X_C, X_B, \operatorname{do}(X_D)) \circ P(X_{F_0}, X_{F_1}, X_H|X_C, \operatorname{do}(X_D)) \tag{26}$$
$$\overset{20}{=} P(X_A|X_{F_0}, X_{F_1}, X_H, X_C, X_B, \operatorname{do}(X_D)) \circ P(X_{F_0}, X_{F_1}, X_H|X_C, \operatorname{do}(X_D)) \tag{27}$$
$$\overset{F=F_0 \dot\cup F_1}{=} P(X_A|X_F, X_H, X_C, X_B, \operatorname{do}(X_D)) \circ P(X_F, X_H|X_C, \operatorname{do}(X_D)) \tag{28}$$
$$= P(X_A|X_F, X_H, X_C, X_B, \operatorname{do}(X_D)) \circ (P(X_H|X_F, X_C, \operatorname{do}(X_D)) \otimes P(X_F|X_C, \operatorname{do}(X_D))) \tag{29}$$
$$\overset{21}{=} P(X_A|X_F, X_H, X_C, X_B, \operatorname{do}(X_D)) \circ (P(X_H|X_F, X_C, X_B, \operatorname{do}(X_D)) \otimes P(X_F|X_C, \operatorname{do}(X_D))) \tag{30}$$
$$= P(X_A|X_F, X_C, X_B, \operatorname{do}(X_D)) \circ P(X_F|X_C, \operatorname{do}(X_D)) \tag{31}$$
$$= P(X_A|X_B, X_F, X_C, \operatorname{do}(X_D)) \circ P(X_F|X_C, \operatorname{do}(X_D)). \tag{32}$$

□

Corollary 4.4.4 (Conditional backdoor covariate adjustment formula)**.** *Consider an IL-BCN:*

$$M = \left(G^+ = (J, V, U, E^+), \quad \left(P_v\left(X_v | \operatorname{do}(X_{\mathrm{Pa}^{G^+}(v)})\right)\right)_{v \in V \cup U}\right).$$

Assume that the conditional backdoor criterion *in the graphs* $G^+_{\operatorname{do}(I_B, D)}$ *holds:*

1. $F \underset{G_{\operatorname{do}(I_B,D)}}{\overset{d}{\perp}} I_B \,|(C \cup D)$, *and:*

2. $A \underset{G_{\operatorname{do}(I_B,D)}}{\overset{d}{\perp}} I_B \,|(B \cup F \cup C \cup D)$.

Then we have the adjustment formula:

$$P(X_A | X_C, \operatorname{do}(X_B, X_D)) = P(X_A | X_B, X_F, X_C, \operatorname{do}(X_D)) \circ P(X_F | X_C, \operatorname{do}(X_D)) \qquad a.s.$$

Proof. It follows directly from theorem 4.4.3 by using $F_1 = H = \emptyset$.
We give another proof in more details here anyways.
By the assumption:

$$A \underset{G_{\operatorname{do}(I_B,D)}}{\overset{d}{\perp}} I_B \,|(B \cup F \cup C \cup D),$$

there exists a Markov kernel that is simultaneously a version of:

$$P(X_A | X_F, X_C, \operatorname{do}(X_B, X_D)) \qquad \text{and} \qquad P(X_A | X_F, X_C, X_B, \operatorname{do}(X_D)).$$

We denote it by:

$$P(X_A | X_F, X_C, \cancel{\operatorname{do}}(X_B), \operatorname{do}(X_D)).$$

By assumption:

$$F \underset{G_{\operatorname{do}(I_B,D)}}{\overset{d}{\perp}} I_B \,|(C \cup D),$$

there exists a Markov kernel that is simultaneously a version of:

$$P(X_F | X_C, \operatorname{do}(X_B, X_D)) \qquad \text{and} \qquad P(X_F | X_C, \operatorname{do}(X_D)).$$

We denote it by:

$$P(X_F | X_C, \cancel{\operatorname{do}(X_B)}, \operatorname{do}(X_D)).$$

Then:

$$P(X_A | X_F, X_C, \cancel{\operatorname{do}}(X_B), \operatorname{do}(X_D)) \circ P(X_F | X_C, \cancel{\operatorname{do}(X_B)}, \operatorname{do}(X_D))$$

is a version of:

$$P(X_A | X_C, \operatorname{do}(X_B, X_D)).$$

□

Corollary 4.4.5 (Backdoor covariate adjustment)**.** *Let the situation be like in theorem 4.4.4 with* $C = D = J = \emptyset$. *Assume that the* backdoor criterion *holds:*

1. $F \overset{d}{\underset{G_{\mathrm{do}(I_B)}}{\perp}} I_B$, *and:*

2. $A \overset{d}{\underset{G_{\mathrm{do}(I_B)}}{\perp}} I_B \,|(B \cup F)$.

Then we have the adjustment formula:

$$P(X_A|\,\mathrm{do}(X_B)) = P(X_A|X_B, X_F) \circ P(X_F) \qquad a.s.$$

Remark 4.4.6 (Tian's ID algorithm)**.** *There exists an algorithm that takes as input the observational CADMG G of an IL-CBN that have densities/mass functions, and sets: $A, B, C \subseteq J \cup V$. It outputs either an algebraic formula for $P(X_A|X_B, \mathrm{do}(X_C))$ in terms of conditional marginals of the observational Markov kernel $P(X_V|\,\mathrm{do}(X_J))$ if the above causal effect is identifiable, or: it outputs FAIL if the above causal effect is not identifiable. The algorithm only uses the do-calculus rules.*

JM: Patrick suggested that the scope of this claim may need to be restricted somehow.

5. Structural Causal Models

Structural Causal Models (SCMs), also known as Structural Equation Models (SEMs), provide a very general class of causal models. They trace back to the early work on path analysis by geneticist Sewall Wright (1921), made their way to econometrics (Strotz & Wold, 1960), and then became popular in AI due to the work of Judea Pearl (and many others). In these lecture notes, we give a modern treatment inspired by our own research on the matter.

5.1. Deterministic Examples

Before giving a formal definition of SCMs, let us revisit the example regarding chocolate consumption and Nobel prizes and try to come up with quantitative causal models for the data generating process.

For a given country, let us 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 consider the following simplified hypothetical causal models:

(i) N causes C ("Nobel prize winning countries celebrate with massive chocolate consumption"). A tacit but important assumption that underlies this model is that N is not caused by C. Assuming an affine relationship, we can model this with the *structural equation*

$$C = \alpha + \beta N \tag{33}$$

where $\alpha, \beta \in \mathbb{R}$ are two parameters. The convention for such a structural equation is that there must be a single variable on the l.h.s. of the equality sign, which indicates that this is the structural equation corresponding to that variable. In the absence of causal cycles, the value of the variable on the l.h.s. is set to the value of the expression on the r.h.s. of the equality sign. While the ordinary ("non-structural") equation $C = \alpha + \beta N$ is usually considered to be equivalent to the equation $N = (C - \alpha)/\beta$ in the sense that their solutions are the same (at least for $\beta \neq 0$), the structural equation $C = \alpha + \beta N$ has an entirely different meaning than the structural equation $N = (C - \alpha)/\beta$ (even for $\beta \neq 0$). We will make this difference more precise when we discuss the causal semantics in terms of interventions in the next section. For now, we will just note that in the absence of causal cycles, the effect is on the l.h.s. of the structural equation, and its direct causes appear on the r.h.s..

In this model, the variable C is *endogenous*, i.e., its structural equation (33) describes how its value is determined by other variables in the model (only N in this case) and model parameters (α and β in this case). The variable N is *exogenous*, since its value is determined by something external to our model: the model remains silent about what determines the value of N.

The model does make two tacit important assumptions about how the values of N are determined that are not apparent from equation (33). These are:

1) the exogenous variable N is *not caused by* the endogenous variable C, and

2) there is *no common cause* of N and C.

The first assumption roughly means that if we were to intervene on chocolate consumption C (e.g., by increasing advertizing for chocolate), this would not lead to a change of N. The second assumption roughly means that any factor (apart from N and C) that causes N does not cause C in any other way than through its effect on N. In particular, this means that the parameters α and β are assumed to not depend on the value of N. In reality, this second assumption is likely to be violated, as the wealth of a country probably causes both C and N directly.

One concrete way to think about the model (and a very useful analogy) is as a small computer program, for example in the language `R`:

```
N_causes_C <- function(N) {
  C <- alpha + beta * N
  return(C)
}
```

The function takes the value of exogenous variable N as input, applies the structural equation (33) to calculate the value of endogenous variable C, and returns that as its output. The variable N is "exogenous" to the function (its value is set externally), while C is "endogenous" (its value is set internally). The computer program analogy also shows that changing the value of C by adapting the code inside the function body does not lead to a change in N.

(ii) C causes N ("chocolate contains brain enhancing chemicals"). We can model this with the structural equation

$$N = \gamma + \delta C \tag{34}$$

with two parameters $\gamma, \delta \in \mathbb{R}$. For this model, we again distinguish exogenous and endogenous variables, but now C is considered the exogenous variable, and N the endogenous variable. A computer program representation of this model in `R` is:

```
C_causes_N <- function(C) {
  N <- gamma + delta * C
  return(N)
}
```

The tacit assumptions are that N does not cause C, and that N and C are not confounded.

(iii) C and N have a common cause, wealth W ("The wealth of a country determines both how much money goes to science and also how much people can spend on chocolate"). We could model this using two structural equations:

$$C = \alpha + \beta W \tag{35}$$

$$N = \gamma + \delta W \tag{36}$$

In this case, we treat the variables C and N both as endogenous, and we treat W as exogenous. The following tacit assumptions are made:

1) W is not caused by C, nor by N;

2) none of the variables is confounded, i.e., there is no other variable that is a common cause of any pair of $\{N, C, W\}$, or even the triple (N, C, W).

Another way to formulate the second assumption is that the values of W, the parameters (α, β) and the parameters (γ, δ) are all determined "independently".

A computer program representation of this model in `R` is:

```
W_causes_C_and_N <- function(W) {
  C <- alpha + beta * W
  N <- gamma + delta * W
  return(c(C,N))
}
```

It takes as input the value of the exogenous variable W, applies the structural equations to calculate the values of the endogenous variables C and N accordingly, and returns a tuple with both values as output.

Even with only two (or three) variables, many different causal models can be made, and the above three illustrate just a small subset of those. Indeed, we could for example combine the third model with a direct causal effect of C on N (or the other way around), and we could assume more complicated confounding relationships involving latent confounders between the variables. In case you wonder how to model the other causal hypotheses (cyclic relationship, selection bias, and functional constraints): we will postpone discussion of these since they are inherently more complicated.

5.2. Hard Interventions

So far, we have been rather informal about the causal meaning of the models (i), (ii) and (iii), which provide different causal explanations for the observed correlation between C and N. We will now make this more precise in terms of hypothetical *interventions* that we could perform on the "system" (a country) consisting of the two variables C and N.

Let us start again with discussing the first model (N causes C), with structural equation (33). We consider two interventions:

- What would happen if Putin forced the Russians to eat c kilograms of chocolate every year? We can model this *intervention on (the structural equation for)* C by modifying structural equation (33) into the following intervened structural equation:

$$C = c. \tag{37}$$

This intervention would obviously lead to a change of the value of C (for most choices of c, i.e., except if c happens to equal $\alpha + \beta N$). It would, however, have no effect on the number of Russian Nobel prize winners: the value of N would remain unaltered, since exogenous variables (in this model, N) are assumed not to be caused by endogenous ones (for this model, C).

Following Pearl, we will denote this intervention as $\text{do}(C = c)$. It is often called a *hard* (or "*surgical*", or "*atomic*", or "*perfect*") intervention because it completely overrides the default causal mechanism that normally determines the value of C expressed by structural equation (33).

The computer program equivalent of this intervention is to replace the line of code `C <- alpha + beta * N` with the assignment `C <- c`. If we consider c to be a (constant) parameter, we get the following "intervened" program:

```
N_causes_C.do_c <- function(N) {
  C <- c
  return(C)
}
```

- What would happen if we somehow enforced that the Russians win a certain number of Nobel prizes (e.g., by manipulating the Nobel prize committee members)? In this case, the causal mechanism determining the chocolate consumption modeled by structural equation (33) would still be in place. Therefore, changing the exogenous value N would generically (i.e., if $\beta \neq 0$) lead to a change of C. In other words, if before the intervention the chocolate consumption is $c = \alpha + \beta n$, and we intervene to set $N = n'$, then after the intervention the value of C will become $c' = \alpha + \beta n'$. This hard intervention is denoted as $\text{do}(N = n')$. It leaves the structural equation (33) invariant. However, to reflect that the value of N is now set to n', we will add the structural equation

$$N = n'$$

to the model, and from now on will treat N as endogenous to reflect that its value has been set. Thus, the intervened model has structural equations:

$$\begin{aligned} N &= n' \\ C &= \alpha + \beta N \end{aligned}$$

and treats both C and N as endogenous variables.

We summarize this discussion in terms of a larger computer program example (in Figure 4) that illustrates what this causal model predicts under different interventions.

Exercise 5.2.1. *Write an analogous* `R` *program for the second model, C causes N. Remember that it treats C as exogenous, N as endogenous, and has structural equation (34). Simulate from the model in the observational setting, after the hard intervention* $\text{do}(C = c)$ *and after the hard intervention* $\text{do}(N = n)$.

For what is presumably the most realistic model (C and N are caused by W) with structural equations (35) and where W is treated as exogenous, and C and N as endogenous, we discuss three hard interventions:

```
### observational model
alpha <- 2.0
beta <- 0.5
N_causes_C <- function(N) {
  C <- alpha + beta * N
  return(C)
}
# query model for the Dutch (n=11):
N_causes_C(11)

### hard intervention: do (C=c)
c <- 12 # observed value for Switzerland
N_causes_C.do_c <- function(N) {
  C <- c
  return(C)
}
# query model for the Dutch (n=11):
N_causes_C.do_c(11)

### hard intervention: do (N=n)
n <- 32 # observed value for Switzerland
N_causes_C.do_n <- function() {
  N <- n
  C <- alpha + beta * N
  return(c(C,N))
}
# query model:
N_causes_C.do_n()
```

Figure 4: Computer program illustrating how to simulate from the causal model "N causes C" under various interventions.

- do($C = c$). This changes the structural equations into:

$$C = c \tag{38}$$
$$N = \gamma + \delta W \tag{39}$$

  Note that while the structural equation for C is changed by the intervention, the structural equation for N is not (i.e., it remains invariant). This reflects the "perfect" character of the intervention: it only changes the structural equation(s) of the intervention target(s) (only C in this case), without any side effects on other structural equations.

- do($N = n$). Similarly, this only changes the structural equation for N, without modifying the structural equation for C, and yields the following intervened structural equations:

$$C = \alpha + \beta W \tag{40}$$
$$N = n \tag{41}$$

- do($W = w$). Since W is exogenous, the structural equations for C and N remain the same as in the "observational" (pre-interventional) setting, i.e., as in (35), and we add a structural equation for W:

$$C = \alpha + \beta W$$
$$N = \gamma + \delta W$$
$$W = w$$

  In this intervened model, all three variables are treated as endogenous variables.

Exercise 5.2.2. *For each of the three hard interventions considered above: what variable values are changed due to the intervention (comparing with the pre-interventional "observational" model)? How do these correspond to the intuitive causal effects between the three variables?*

5.3. Soft Interventions

A more general class of interventions are *soft interventions*. These modify the expression on the r.h.s. of a structural equation, but without adding new variables to that expression. Special cases are hard interventions on endogenous variables.

As an example, consider the model with the confounder W for C and N, with structural equations:

$$C = \alpha + \beta W$$
$$N = \gamma + \delta W.$$

An example of a soft intervention on C is a change in the parameter β, replacing it with β' (perhaps to reflect a change in the chocolate price):

$$C = \alpha + \beta' W$$
$$N = \gamma + \delta W.$$

Another example would be to change the functional form of the structural equation:

$$C = \beta\gamma + \alpha W^3$$
$$N = \gamma + \delta W.$$

In contrast to a hard intervention on C, which would remove all occurrences of endogenous and exogenous variables at the r.h.s. of the structural equation for C, this soft intervention only changes the functional / parametric form of the equation, but its value may still depend on the variables that appeared in the original structural equation.

In general, a soft intervention on C will result in the following structural equations:

$$C = f_\theta(W)$$
$$N = \gamma + \delta W$$

for some function $f_\theta : \mathbb{R} \to \mathbb{R}$ parameterized by some parameter $\theta \in \Theta$. Similarly, a soft intervention on N will result in

$$C = \alpha + \beta W$$
$$N = g_\theta(W)$$

for some function $g_\theta : \mathbb{R} \to \mathbb{R}$ parameterized by some parameter $\theta \in \Theta$.

5.4. Intervention Variables

It can be very convenient to introduce explicit variables that indicate whether or how interventions have been performed. These are often referred to as *intervention variables* (and are special cases of so-called *context variables*, *environment variables* and *domain indicators*).

Often, intervention variables can be considered as exogenous. For instance, if their values (i.e., which intervention is performed, and how) are determined earlier in time than the values of the endogenous variables, which means that the endogenous variables cannot cause the intervention variables (except, perhaps, if time travel becomes possible). An example is a scientific experiment in which the experimenter prepares the system in some experimental condition before performing measurements, which implies that the experimental condition cannot be influenced by the measured state of the system. A concrete instance of such an intervention variable is the coin flip that determines whether a subject enters the treatment or control group in a randomized controlled trial. However, if a doctor prescribes a certain treatment based on the state of the patient determined through a diagnosis, then the treatment variable is influenced by the

patient's state and should therefore not be considered to be exogenous with respect to the patient's state.

Let us discuss the randomized controlled trial example as described in Principle 1.3.1 in more detail. Let X be the variable that describes treatment, and Y be the variable that describes the outcome. Introduce the binary intervention variable C that indicates whether the subject enters the test group ($C = 1$) or the control group ($C = 0$). In practice, one uses a coin or, nowadays, a random number generator, to determine the value of C. Due to the experimental setup of the RCT, we may therefore safely assume Y and X not to cause C. Thus we can consider C as an exogenous variable, and can write down the following structural equations to model the RCT, treating X and Y as endogenous variables:

$$\begin{aligned} X &= \begin{cases} x_0 & \text{if } C = 0, \\ x_1 & \text{if } C = 1 \end{cases} \\ Y &= \alpha + \beta X. \end{aligned}$$

The parameter β is often referred to as "the causal effect of X on Y", in the sense of being the effect that a unit change of X has on Y. In addition to the assumption that C is not caused by X or Y, an important assumption is also that there is no confounding between C and Y. This boils down to assuming that the outcome of the coin flip is not determined by anything else that causally influences the outcome Y (in particular, there is no room for God's will in this model).

There is no need for an intervention variable to be binary or discrete. An example of a continuous intervention variable is the dose of the drug used for treatment. The dose often quantitatively affects the outcome. This dependence is visualized in dose-response curves that link a quantitative response to the dose of the drug that was administered. The following so-called Hill model is often used:

$$Y = \frac{Y_{max}}{1 + \left(\frac{D}{D_{50}}\right)^{-n}}$$

where Y is the response, D the dose, and Y_{max}, D_{50} and n are parameters (the maximum response, the potency, and the Hill coefficient, respectively). We can consider D to be a continuous (non-negative) intervention variable, that not only indicates whether an intervention was performed ($D > 0$?), but also how the intervention was performed (the administered dose). It is again obvious from the setting that the response does not causally influence the dose. However, if there is no explicit and careful randomization, one cannot always assume that there is no common cause of dose and response. For example, if physicians decide not to treat patients with terminal cancer, then the stage of the cancer is a common cause of dose and response.

5.5. Modeling Uncertainty

So far, we have not introduced any randomness into our causal models. This is often an important step when modeling a system. While this is a *sine qua non* for purely

probabilistic/statistical modeling, it is optional—yet common—in causal modeling.

The main reason for introducing stochasticity into a model is to deal with missing information about the system state. For example, when rolling a die, if one would have access to the exact initial state (position, momentum, angular momentum) of the die once released, one could calculate the outcome of the roll as a (complicated) deterministic function of the initial state. In practice, it is neither feasible to measure or control the initial state of the die after it is released from the hand, nor to calculate this deterministic function. Therefore, instead of attempting to model this meticulously, one can deal with any remaining uncertainty by modeling its probability distribution (which could either describe the subjective beliefs about the value of a variable in a Bayesian interpretation, or, in a frequentist interpretation, a description of its statistics given an infinite ensemble consisting of copies of the system).

In causal modeling, we often have to incorporate stochasticity in this way. Therefore, in a structural causal model, we will specify certain probability distributions on subsets of the *exogenous* variables to explicitly represent "noise" that may affect the values of the endogenous variables. This leads us to distinguish three types of variables: endogenous variables, exogenous input variables, and exogenous random variables. The only difference between the exogenous input variables and the exogenous random variables is that the model specifies the probability distributions of the latter, while not making any assumptions on the values assumed by the former. Often, the exogenous random variables are introduced to model "noise" and are considered latent; hence they are sometimes referred to as "noise" or "disturbance" variables.

As an example, we will again revisit the chocolate consumption and Nobel price winners example. Consider the causal model C causes N, based on the hypothesis that chocolate consumption enhances cognitive abilities. When looking at the data in Figure 1, it is obvious that an exact affine relationship of the form $N = \gamma + \delta C$ cannot be very accurate, as for example The Netherlands has the same chocolate consumption as Australia, but obtained considerably more Nobel prizes per capita. We can hypothesize that other factors determine the value of N, apart from C. Rather than attempting to meticulously model all of these factors, we will introduce just a single "effective" or "aggregated" exogenous random variable E that expresses their residual influence on N. We thus refine our structural equation for N to:

$$N = \gamma + \delta C + E \tag{42}$$

and in addition we assume that E has a certain probability distribution P^E. For concreteness, we could assume that $E \sim \mathcal{N}(0, \sigma^2)$, i.e., that E has a Gaussian distribution with zero mean and variance σ^2, where we introduced a new parameter σ^2 to the model. Summarizing, the stochastic structural causal model then becomes:

$$\begin{aligned} E &\sim \mathcal{N}(0, \sigma^2) \\ N &= \gamma + \delta C + E, \end{aligned}$$

with C an exogenous input variable, E an exogenous random variable, and N an endogenous variable, and parameters γ, δ, σ^2.

The model can again be implemented as a computer program in `R`:

```
C_causes_N_stoch <- function(C) {
  E <- rnorm(1,mean = 0,sd = sigma)
  N <- gamma + delta * C + E
  return(N)
}
```

This function takes as input the value of C, then samples an (independent) random value for E, and uses that to calculate the value for N that is returned. Here, N is the endogenous variable, C is an exogenous input variable, and E is an exogenous random variable.

Note that the analogy is not entirely accurate, as the structural causal model prescribes the probability distribution of E, while the computer program samples a value from this distribution. The reason is that a language like R does not allow us to express probability distributions; for that, we would need a probabilistic programming language, but those are less well known and therefore less useful for our didactic purpose. Henceforth we will think of computer implementations of a structural causal model as programs that *sample* from the model, and whose elements correspond directly with those of the model.

5.6. Formal Definitions: Structural Causal Models

After this lengthy motivation, we will now define structural causal models formally. In contrast with many definitions encountered in the literature,[11] we will explicitly distinguish three types of variables: exogenous random variables and endogenous variables, and exogenous input variables. This may appear to complicate matters at first sight, but will instead provide formal clarity later on.

Definition 5.6.1 (Structural Causal Model)**.** *A Structural Causal Model is a tuple* $\langle J, V, W, \mathcal{X}, P, f \rangle$ *with*[12]

- 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 $\mathcal{X} = \prod_{i \in J \cup V \cup W} \mathcal{X}_i$ *is a product of standard measurable spaces* $\mathcal{X}_i$*;*
- *the* exogenous distribution P *is a probability distribution on* $\mathcal{X}_W$ *that factorizes as a product* $P = \bigotimes_{w \in W} P_w$ *of probability distributions* $P_w \in \mathcal{P}(\mathcal{X}_w)$*;*
- *the* causal mechanism *is specified by the measurable function* $f : \mathcal{X} \to \mathcal{X}_V$*.*

[11] For example, [Pea09] only formally distinguishes exogenous random variables and endogenous variables.

[12] 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 SCM 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 $* = \{*\}$ 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, it often happens that there are no exogenous input variables (if J is empty).

Often, the causal mechanism f and the exogenous distribution P depend (in a measurable way) on exogenous parameters *$\theta \in \Theta$, which we may make explicit by writing f_θ and P_θ instead, giving a parameterized SCM $\mathcal{M}_\theta = \langle J, V, W, \mathcal{X}, P_\theta, f_\theta \rangle$. The family $(\mathcal{M}_\theta)_{\theta \in \Theta}$ is then an SCM family.*[13]

One can also think about an SCM as describing an input/output system, with free inputs J, random inputs W with distribution P, outputs V and input/output mechanism f. Note that a statistical model can be seen as a special case of a SCM that only has exogenous random variables ($J = V = \emptyset$). A complementary subclass is formed by deterministic models without exogenous random variables ($W = \emptyset$) but with exogenous input variables and endogenous variables that we discussed in Section 5.1. In this sense, structural causal models can be regarded as a marriage of statistical models as traditionally used in statistics with deterministic causal models that are used informally in disciplines like physics and engineering.

Remark 5.6.2. *There are three crucial assumptions embodied in the modeling approach using SCMs:*

1. *The distinction between endogenous and exogenous variables: exogenous variables (i.e., exogenous input variables, exogenous random variables, and exogenous parameters) are not caused by endogenous variables, by assumption;*

2. *Exogenous random variables are mutually independent, and independent of the exogenous input variables in the sense that their probability distribution does not depend on the joint value of all exogenous input variables; however, we do allow for "dependencies" between exogenous input variables;*

3. *Exogenous parameters are distinguished from exogenous random variables in that the former describe "population" properties whereas the latter describe "individual" quantities.*

Often (but not always) the exogenous random variables are latent (i.e., not observed), and are used as "noise" variables to incorporate remaining uncertainty in the values of the endogenous variables. However, since it may depend on the context whether a variable is observed or latent (e.g., in a training data set, the prediction target is typically observed, whereas it is latent in a test data set), we will not incorporate formal assumptions regarding which variables are observed and which are latent into the model. In this point, our exposition deviates from many accounts on SCMs in the literature.

In our informal motivation, models were defined by structural equations (with specified probability distributions on exogenous random variables). These structural equations are not explicitly included in Definition 5.6.1 but can be derived from the causal mechanism as the equations that determine the solutions of the model. This is done as follows.

[13] In line with the convention in ML, the word "model" refers to a SCM 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".

Definition 5.6.3 (Solutions and outcomes for a specified input)**.** *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ be an SCM 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* solution of $\mathcal{M}$ for input x_J *if the following two conditions hold:*

1. *its W-component has the exogenous distribution specified by $\mathcal{M}$:*

$$X_W^{\mathrm{do}(x_J)} \sim P,$$

2. *it satisfies the structural equations entailed by $\mathcal{M}$ for input x_J:*

$$X_V^{\mathrm{do}(x_J)} = f\big(x_J, X_V^{\mathrm{do}(x_J)}, X_W^{\mathrm{do}(x_J)}\big) \ \text{a.s.}. \tag{43}$$

When considering only the subset $O \subseteq V \cup W$ to be observed (and $(V \cup W) \setminus O$ latent) we call $X_O^{\mathrm{do}(x_J)} := (X_{O \cap V}^{\mathrm{do}(x_J)}, X_{O \cap W}^{\mathrm{do}(x_J)})$ a potential outcome of $\mathcal{M}$ for input x_J.[14]

In practice, an SCM is often specified more informally by writing down the corresponding structural equations (43) 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 distribution specified, are then implicitly taken as exogenous inputs or parameters.

The following remark relates the terminology to the cases most often considered in the literature.

Remark 5.6.4. *For the special case of an SCM with no exogenous input variables ($J = \emptyset$), we can drop the superscript "$\mathrm{do}(x_J)$" and the "for input x_J". This gives a formulation more commonly encountered in the literature: A random variable $X_{V \cup W}$ with codomain $\mathcal{X}_V \times \mathcal{X}_W$ is called a* solution of $\mathcal{M}$ *if $X_W \sim P$ and it satisfies the structural equation:*

$$X_v = f_v\big(X_V, X_W\big) \ \text{a.s.} \tag{44}$$

for each $v \in V$. When considering only the subset $O \subseteq V \cup W$ to be observed (and $(V \cup W) \setminus O$ latent) we call $X_O := (X_{O \cap V}, X_{O \cap W})$ an outcome of $\mathcal{M}$*. A common convention is to take W as latent (i.e., $O = V$), and under that convention an outcome of $\mathcal{M}$ is a random variable with codomain $\mathcal{X}_V$ that satisfies the structural equations (44) for some $X_W \sim P$.*

We often encounter families of solutions for specified inputs that are also measurable w.r.t. the exogenous input.

[14] Note that when considering two potential outcomes $X_O^{\mathrm{do}(x_J)}, X_O^{\mathrm{do}(x'_J)}$ of $\mathcal{M}$ for different inputs x_J, x'_J we do *not* necessarily assume that the two random variables $X_W^{\mathrm{do}(x_J)}$ and $X_W^{\mathrm{do}(x'_J)}$ are the same; we only assume that they have the same distribution. This choice has been made to avoid introducing implicitly defined counterfactuals. In our opinion, it is better to explicitly introduce these with the twinning construction, as will be described in Section 6.2, because this enforces one to think about which variables are shared across potential worlds and which are copied (resampled).

Definition 5.6.5 (Solutions of an SCM)**.** *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be an SCM and $(\Omega, \Sigma, \mathbb{P})$ a probability space. A family of solutions $(X_{V\cup W}^{\mathrm{do}(x_J)})_{x_J\in\mathcal{X}_J}$ of $\mathcal{M}$, one for each input x_J, with common domain Ω, is called a* solution of $\mathcal{M}$ *if the mapping*

$$\Omega \times \mathcal{X}_J \to \mathcal{X}_V \times \mathcal{X}_W : (\omega, x_J) \mapsto \big(X_V^{\mathrm{do}(x_J)}(\omega), X_W^{\mathrm{do}(x_J)}(\omega)\big)$$

is measurable.

Not every SCM has solutions, since not every set of equations has a solution. Also, if they exist, solutions are not necessarily unique.

Example 5.6.6. *Consider an SCM with parameters α, β, c, endogenous real variables X_1, X_2 and structural equations*

$$\begin{cases} X_1 = \alpha X_2 \\ X_2 = \beta X_1 + c \end{cases}$$

If $\alpha\beta = 1$ and $c = 0$, then $(X_1, X_2) = (\alpha x, x)$ is a solution, no matter which value $x \in \mathbb{R}$ we choose. If $\alpha\beta = 1$ and $c \neq 0$, then there is no solution. If $\alpha\beta \neq 1$, then the solution is unique: $(X_1, X_2) = (\frac{\alpha c}{1-\alpha\beta}, \frac{c}{1-\alpha\beta})$.

We can construct solutions and potential outcomes in terms of solution functions of an SCM:

Definition 5.6.7 (Solution function of an SCM)**.** *Given an SCM $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$, we call a measurable function $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ a* solution function of $\mathcal{M}$ *if $g(x_J, x_W)$ satisfies the structural equations for all $x_W \in \mathcal{X}_W$, for all $x_J \in \mathcal{X}_J$:*

$$g(x_J, x_W) = f\big(x_J, g(x_J, x_W), x_W\big).$$

Remark 5.6.8. *If g is a solution function for SCM $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$, then for any random variable $X_W \sim P$:*

1. *$(g(x_J, X_W), X_W)_{x_J\in X_J}$ is a solution of $\mathcal{M}$, and*
2. *$X_O^{\mathrm{do}(x_J)} := (g_O(x_J, X_W), X_{O\cap W})$ is a (potential) outcome for $\mathcal{M}$ for input x_J (where $O \subseteq V \cup W$ are considered observed).*

Not all (potential) outcomes or solutions 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 5.6.9. *For an SCM 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 (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 as the push-forward through a solution function.

Example 5.6.10. *We formalize the three models we discussed in Section 5.1 (and a variant) as SCMs.*

- *N causes C: $J = \{N\}$, $V = \{C\}$, $W = \emptyset$, $\theta = (\alpha, \beta) \in \mathbb{R}^2$, $\mathcal{X}_N = \mathbb{R}$, $\mathcal{X}_C = \mathbb{R}$, $f : (x_N, x_C) \mapsto \alpha + \beta x_N$. It has a unique solution function, $g : x_N \mapsto \alpha + \beta x_N$. It has unique outcomes of the form $X_C^{\mathrm{do}(x_N)} = \alpha + \beta x_N$.*
- *C causes N: $J = \{C\}$, $V = \{N\}$, $W = \emptyset$, $\theta = (\gamma, \delta) \in \mathbb{R}^2$, $\mathcal{X}_C = \mathbb{R}$, $\mathcal{X}_N = \mathbb{R}$, $f : (x_C, x_N) \mapsto \gamma + \delta x_C$. It has solution function, $g : x_C \mapsto \gamma + \delta x_C$. It has outcomes of the form $X_N^{\mathrm{do}(x_C)} = \gamma + \delta x_C$.*
- *W causes C and N: $J = \{W\}$, $V = \{C, N\}$, $W = \emptyset$, $\theta = (\alpha, \beta, \gamma, \delta) \in \mathbb{R}^4$, $\mathcal{X}_W = \mathbb{R}$, $\mathcal{X}_C = \mathbb{R}$, $\mathcal{X}_N = \mathbb{R}$, $f : (x_W, x_C, x_N) \mapsto (\alpha + \beta x_W, \gamma + \delta x_W) \in \mathcal{X}_C \times \mathcal{X}_N$. It has solution function, $g : x_W \mapsto (\alpha + \beta x_W, \gamma + \delta x_W) \in \mathcal{X}_C \times \mathcal{X}_N$. It has outcomes of the form $X_V^{\mathrm{do}(x_W)} = (\alpha + \beta x_W, \gamma + \delta x_W)$.*
- *W causes C and N, stochastic version: $J = \emptyset$, $V = \{C, N\}$, $W = \{W\}$, $\theta = (\alpha, \beta, \gamma, \delta, \sigma) \in \mathbb{R}^4 \times [0, \infty)$, $\mathcal{X}_W = \mathbb{R}$, $\mathcal{X}_C = \mathbb{R}$, $\mathcal{X}_N = \mathbb{R}$, $f : (x_C, x_N, x_W) \mapsto (\alpha + \beta x_W, \gamma + \delta x_W) \in \mathcal{X}_C \times \mathcal{X}_N$, $P = \mathcal{N}(0, \sigma^2)$. It has solution function $g : x_W \mapsto (\alpha + \beta x_W, \gamma + \delta x_W) \in \mathcal{X}_C \times \mathcal{X}_N$. Its outcomes are essentially of the form $X_V = (\alpha + \beta X_W, \gamma + \delta X_W)$ for any random variable $X_W \sim \mathcal{N}(0, \sigma^2)$.*

The reason that equations (43) 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 should become clear after the next section.

5.7. Formal Definitions: Interventions

In this section we define interventions as operations on SCMs that map a given SCM and an intervention target (and optionally, an intervention value or distribution) to an intervened SCM. The operation may change the variable types. 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.

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 5.7.1 (Hard intervention with specified target values)**.** *Given an SCM* $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$*, an intervention target* $T \subseteq J \cup V \cup W$ *and an intervention value* $\xi_T \in \mathcal{X}_T$*, we define the* intervened SCM $\mathcal{M}_{\mathrm{do}(X_T = \xi_T)}$ *as the tuple* $\langle J \setminus T, V \cup T, W \setminus T, \mathcal{X}, P_{W \setminus T}, (f_{V \setminus T}, \xi_T) \rangle$*. 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 $j \in V \cup T$*, and the intervened exogenous distribution is obtained by marginalizing:*

$$P_{W \setminus T} = \bigotimes_{w \in W \setminus T} P_w.$$

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 $X_j^{\mathrm{do}(x_J)} = f_j(x_J, X_V^{\mathrm{do}(x_J)}, X_W)$ 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 and exogenous random 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 5.7.2. *A hard intervention* $\mathrm{do}(X_N = \xi_N)$ *changes the stochastic SCM "W causes C and N" from Example 5.6.10 into:* $J = \emptyset$*,* $V = \{C, N\}$*,* $W = \{W\}$*,* $\theta = (\alpha, \beta, \gamma, \delta, \sigma) \in \mathbb{R}^4 \times [0, \infty)$*,* $\mathcal{X}_W = \mathbb{R}$*,* $\mathcal{X}_C = \mathbb{R}$*,* $\mathcal{X}_N = \mathbb{R}$*,* $\tilde{f} : (x_C, x_N, x_W) \mapsto (\alpha + \beta x_W, \xi_N) \in \mathcal{X}_C \times \mathcal{X}_N$*,* $P = \mathcal{N}(0, \sigma^2)$*. One of its solution functions is* $\tilde{g} : x_W \mapsto (\alpha + \beta x_W, \xi_N)$*.*

Another common variant of hard interventions are stochastic hard interventions, where the intervention values are drawn independently from a specified (independent) distribution.

Definition 5.7.3 (Stochastic hard intervention)**.** *Given an SCM* $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$*, an intervention target* $T \subseteq J \cup V \cup W$ *and an intervention target distribution* $Q_T \in \bigotimes_{t \in T} \mathcal{P}(\mathcal{X}_t)$*, we define the* intervened SCM $\mathcal{M}_{\mathrm{do}(X_T \sim Q_T)}$ *as the SCM* $\langle J \setminus T, V \setminus T, W \cup T, \mathcal{X}, P_{W \setminus T} \otimes Q_T, f_{V \setminus T} \rangle$*. More explicitly, the intervened exogenous distribution is given by*

$$P_{W \setminus T} \otimes Q_T = \left[\bigotimes_{w \in W \setminus T} P_w \right] \otimes \left[\bigotimes_{t \in T} Q_t \right].$$

Intuitively, this assigns random values to the intervention target variables by sampling from an independent and factorizing intervention distribution Q_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 5.7.4. *The stochastic SCM "W causes C and N" from Example 5.6.10 is obtained from a stochastic intervention on the non-stochastic SCM "W causes C and N" from that example.*

The third variant of hard interventions only specifies the intervention targets, but makes no assertions about the intervention values (not even its distribution).

Definition 5.7.5 (Hard intervention with unspecified value)**.** *Given an SCM $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ and an intervention target $T \subseteq J \cup V \cup W$, we define the* intervened SCM $\mathcal{M}_{\mathrm{do}(T)}$ *as the SCM $\langle J \cup T, V \setminus T, W \setminus T, \mathcal{X}, P_{W \setminus T}, f_{V \setminus T} \rangle$.*

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 5.7.6. *A hard intervention $\mathrm{do}(N)$ on the stochastic SCM "W causes C and N" from Example 5.6.10 yields an intervened SCM with: $J = \{N\}$, $V = \{C\}$, $W = \{W\}$, $\theta = (\alpha, \beta, \gamma, \delta, \sigma) \in \mathbb{R}^4 \times [0, \infty)$, $\mathcal{X}_W = \mathbb{R}$, $\mathcal{X}_C = \mathbb{R}$, $\mathcal{X}_N = \mathbb{R}$, $f_C : (x_N, x_C, x_W) \mapsto \alpha + \beta x_W$, $P = \mathcal{N}(0, \sigma^2)$. It has solution function $g : (x_N, x_W) \mapsto \alpha + \beta x_W$. It has outcomes essentially of the form $X_C^{\mathrm{do}(x_N)} = \alpha + \beta X_W$ for $X_W \sim \mathcal{N}(0, \sigma^2)$ that do not depend on the value x_N used for the hard intervention on N, reflecting that X_N does not cause X_C. On the other hand, the outcomes $X_N^{\mathrm{do}(\xi_N)} = \xi_N$ reflect that a hard intervention on N does affect the value of X_N (unsurprisingly).*

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 mechanism so that the values of the intervened endogenous variables are no longer determined by other endogenous variables, but 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. Since most accounts on SCMs do not offer exogenous input variables, the two variants most often seen in the literature are the latter two.

Proposition 5.7.7. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ be an SCM. Hard interventions $\mathrm{do}(T_1 \dots)$, $\mathrm{do}(T_2 \dots)$ with disjoint targets $T_1, T_2 \subseteq J \cup W \cup V$ (of any of the three variants) commute:*

$$(\mathcal{M}_{\mathrm{do}(T_1 \dots)})_{\mathrm{do}(T_2 \dots)} = (\mathcal{M}_{\mathrm{do}(T_2 \dots)})_{\mathrm{do}(T_1 \dots)}.$$

Proof. This follows by writing out the definitions and checking commutativity of the operations performed on the various components of the SCM tuple one-by-one. □

Finally, we will briefly discuss how soft interventions can be formally modeled. While perhaps the simplest way would be to replace the targeted components f_j of the causal mechanism by other functions $\tilde{f}_j$, the use of intervention variables as discussed in Section 5.4 is often very convenient to model the different interventional contexts in a single model.

6. Structural Causal Models

6.1. Unique solvability

Definition 6.1.1. *An SCM $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ is called* uniquely solvable *if it has a unique solution function. This means two things: (i) it has a solution function $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$, i.e., a measurable function that satisfies*

$$\forall x_W \in \mathcal{X}_W \forall x_J \in \mathcal{X}_J : \quad g(x_J, x_W) = f\big(x_J, g(x_J, x_W), x_W\big);$$

(ii) all its solution functions must equal g, i.e., for any solution function $\tilde{g} : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$, we have that for all $x_W \in \mathcal{X}_W$, for all $x_J \in \mathcal{X}_J$:

$$g(x_J, x_W) = \tilde{g}(x_J, x_W).$$

The following result gives two useful properties of unique solvability. The first provides an equivalent formulation that makes it easier to check whether an SCM is uniquely solvable, the second provides an important consequence regarding the Markov kernels of the SCM.

Theorem 6.1.2. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be an SCM.*

1. *$\mathcal{M}$ is uniquely solvable if and only if for all $x_W \in \mathcal{X}_W$, for all $x_J \in \mathcal{X}_J$, the equation*

$$x_V = f(x_J, x_V, x_W)$$

has a unique solution for $x_V \in \mathcal{X}_V$.

2. *If $\mathcal{M}$ is uniquely solvable, it has a unique Markov kernel*

$$P_{\mathcal{M}}(X_V, X_W \mid \mathrm{do}(X_J)) = (g, \mathrm{id}_{\mathcal{X}_W})_* \bigotimes_{w \in W} P_w.$$

Proof. 1. " $\Longrightarrow$ ": Let $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ be a solution function for $\mathcal{M}$. Let $\xi_W \in \mathcal{X}_W$, $\xi_J \in \mathcal{X}_J$. The equation $x_V = f(x_V, \xi_J, \xi_W)$ does have a solution for x_V (indeed, $g(\xi_J, \xi_W)$ is such a solution). Suppose its solution is not unique, i.e., it has another solution $\tilde{x}_v \neq g(\xi_J, \xi_W)$. Consider the modified function

$$\tilde{g} : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V : (x_J, x_W) \mapsto \begin{cases} \tilde{x}_V & x_J = \xi_J \wedge x_W = \xi_W \\ g(x_J, x_W) & \text{otherwise} \end{cases}$$

$\tilde{g}$ is measurable (because g is measurable), and hence it provides a solution function. However, $\tilde{g} \neq g$, which contradicts the assumed unique solvability.

" $\Longleftarrow$ ": This boils down to proving that the measurability of f and the uniqueness of the solutions implies the measurability of the solution function. We exploit that we are dealing with standard measurable spaces. Define the function g :

$\mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ by letting $g(x_J, x_W)$ be the (unique) solution x_V of the equation $x_V = f(x_V, x_J, x_W)$, with $x_V \in \mathcal{X}_V$. The graph of this function is

$$\operatorname{graph}(g) = \{(x_J, x_W, x_V) \in \mathcal{X}_J \times \mathcal{X}_W \times \mathcal{X}_V : x_V = g(x_J, x_W)\}.$$

By assumption, we have that $x_V = g(x_J, x_W) \iff x_V = f(x_V, x_J, x_W)$ for all $x \in \mathcal{X}$. Hence,

$$\operatorname{graph}(g) = \{(x_J, x_W, x_V) \in \mathcal{X}_J \times \mathcal{X}_W \times \mathcal{X}_V : x_V = f(x_V, x_J, x_W)\}.$$

Defining the function $h : \mathcal{X}_J \times \mathcal{X}_W \times \mathcal{X}_V \to \mathcal{X}_V \times \mathcal{X}_V : (x_J, x_W, x_V) \mapsto (x_V, f(x_V, x_J, x_W))$ and the diagonal $\Delta = \{(x_V, x_V) : x_V \in \mathcal{X}_V\} \subseteq \mathcal{X}_V^2$, this shows that

$$\operatorname{graph}(g) = h^{-1}(\Delta).$$

Since h is measurable and Δ is a measurable set (because $\mathcal{X}_V$ is Hausdorff), $h^{-1}(\Delta)$ is a measurable set. By [Kec95, 14.12], because all spaces are (isomorphic to) Borel spaces, the fact that $\operatorname{graph}(g)$ is a measurable set implies that g is a measurable function. Hence g is a solution function. The unique solvability of $\mathcal{M}$ follows since this is the only possible solution function.

2. Assume $\mathcal{M}$ to be uniquely solvable and let $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ be its unique solution function. The push-forward $(g, \mathrm{id}_{\mathcal{X}_W})_*(P)$ of the exogenous distribution P of $\mathcal{M}$ (interpreted as a constant Markov kernel from $\mathcal{X}_J$ to $\mathcal{X}_W$) provides a Markov kernel for $\mathcal{M}$.

   Let $K(X_V, X_W \mid X_J)$ denote any Markov kernel for $\mathcal{M}$ obtained from a family $(X_V^{\mathrm{do}(x_J)}, X_W)_{x_J \in \mathcal{X}_J}$ of solutions. Pick any $x_J \in \mathcal{X}_J$. We have to show that $K(X_V, X_W \mid X_J = x_J)$ is a unique distribution. Since $(X_V^{\mathrm{do}(x_J)}, X_W)$ is a solution of $\mathcal{M}$, we have that $X_W \sim P$ and

$$X_V^{\mathrm{do}(x_J)} = f(x_J, X_V^{\mathrm{do}(x_J)}, X_W) \quad \text{a.s.}$$

   This implies that

$$X_V^{\mathrm{do}(x_J)} = g(x_J, X_W) \quad \text{a.s.}$$

   By modifying the random variable $X_V^{\mathrm{do}(x_J)}$ on a null set, we can obtain a random variable $\tilde{X_V}^{\mathrm{do}(x_J)}$ such that we get equality everywhere:

$$\tilde{X_V}^{\mathrm{do}(x_J)} = g(x_J, X_W).$$

   This shows that the distribution of $(\tilde{X_V}^{\mathrm{do}(x_J)}, X_W)$, and hence that of $(X_V^{\mathrm{do}(x_J)}, X_W)$, is that of the push-forward of P through the unique function $g_{x_J} : \mathcal{X}_W \to \mathcal{X}_V : x_W \mapsto (g(x_J, x_W), x_W)$, which is unique.

□

The first equivalence states that one gets the *measurability* of the solution function for free from the measurability of the causal mechanism f and the uniqueness of the solutions of the structural equations. Note that for the special case of no exogenous input variables ($J = \emptyset$), the above shows that uniquely solvable SCMs induce a unique observational distribution.

For SCMs whose causal mechanism are linear in terms of the endogenous variables, we can give a sufficient condition for unique solvability, and explicitly write down the form of their solution function.

Proposition 6.1.3. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ be an SCM such that all its endogenous variables are real-valued (i.e., $\mathcal{X}_v = \mathbb{R}$ for $v \in V$), and each component of the causal mechanism is an affine combination of endogenous variables with coefficients that may depend on exogenous variables, i.e., of the form*

$$f_v(x) = \sum_{u \in V} B_{vu}(x_J, x_W) x_u + c_v(x_J, x_W),$$

where $B(x_J, x_W) \in \mathbb{R}^{V \times V}$ is a family of matrices and $c(x_J, x_W) \in \mathbb{R}^V$ is a family of real-valued offsets.

Then, for any set $L \subseteq V$, $(\mathcal{M})_{\mathrm{do}(V \setminus L)}$ is uniquely solvable if and only if the matrices $I_L - B_{LL}(x_J, x_W)$ are invertible for all $x_J \in \mathcal{X}_J, x_W \in \mathcal{X}_W$, where $I_L \in \mathbb{R}^{L \times L}$ the identity matrix and $B_{LL}(x_J, x_W) \in \mathbb{R}^{L \times L}$ the submatrix of $B(x_J, x_W)$. Its unique solution function is then:

$$\begin{aligned} g : \mathbb{R}^{V \setminus L} \times \mathbb{R}^{J \cup W} &\to \mathbb{R}^L \\ : (x_{V \setminus L}, x_J, x_W) &\mapsto (I_L - B_{LL}(x_J, x_W))^{-1} \big(B_{L, V \setminus L}(x_J, x_W) x_{V \setminus L} + c_L(x_J, x_W) \big). \end{aligned}$$

If an SCM is uniquely solvable, this does not necessarily mean that it is still 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 (mostly) restrict our attention to a subclass of SCMs that we refer to as *simple SCMs*, which are SCMs that are uniquely solvable and remain so after any hard intervention:

Definition 6.1.4. *An SCM $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ is called* simple *if for any $T \subseteq J \cup V \cup W$, the intervened SCM $\mathcal{M}_{\mathrm{do}(T)}$ is uniquely solvable.*

Note that this includes unique solvability of $\mathcal{M}$ itself for $T = \emptyset$.

Example 6.1.5. *Consider an SCM 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:

$$\begin{aligned} X_1 &= W_1 \\ X_2 &= W_2 \\ X_3 &= \frac{W_3 + W_1 W_4}{1 - W_1 W_2} \\ X_4 &= \frac{W_2 W_3 + W_4}{1 - W_2 W_1} \end{aligned}$$

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. 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 SCM is simple.

Theorem 6.1.2 shows that the Markov kernels in the following definition are all uniquely defined.

Definition 6.1.6. *If SCM $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ is simple, and $O \subseteq V \cup W$ is considered observed, then $\mathcal{M}$ induces the* observational Markov kernel $P_{\mathcal{M}}(X_O \mid \mathrm{do}(X_J))$*, and for any intervention target $T \subseteq O$, it induces the* interventional Markov kernel

$$P_{\mathcal{M}}(X_{O \setminus T} \mid \mathrm{do}(X_J), \mathrm{do}(X_T)) := P_{\mathcal{M}_{\mathrm{do}(T)}}(X_{O \setminus T} \mid \mathrm{do}(X_{J \cup T})),$$

as the marginal Markov kernel of the intervened SCM $\mathcal{M}_{\mathrm{do}(T)}$.

The two notations for interventional Markov kernels of simple SCMs will be used interchangeably.

The fact that these SCMs are relatively simple to deal with (because we do not have to worry about non-existence or non-uniqueness of solutions or partial solutions) motivated their name. Even though simplicity is a strong assumption, the class of simple SCMs is more expressive than the class of IL-CBNs for two reasons: (i) it allows for causal cycles, and (ii) it allows to define counterfactuals.

6.2. Counterfactuals through twinning

Counterfactuals are hypothetical statements (or questions) regarding the effects of some change that is contrary-to-fact. 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 change (e.g., since you were in a healthy state and did not drink alcohol, you most likely will feel well in this alternative world).[15]

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 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.[16] It mimics the reasoning steps we mentally perform when thinking about counterfactuals by constructing an "actual" ("factual") 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 6.2.1. *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 key definition is:

[15] The word "counterfactual" is also often used in the causality literature as a synonym for potential outcomes, which are not necessarily contrary-to-fact. This would correspond with "If you drink too much beer you will get a hangover." It is important to be aware of this to avoid possible confusion.

[16] Consider this a warning before attempting to predict counterfactual statements in a data-driven way, for example, using a neural network.

Definition 6.2.2. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be an SCM. We define the* twinned SCM *of $\mathcal{M}$ as the SCM $\mathcal{M}^{\text{twin}} = \langle J \dot\cup J', V \dot\cup V', W, \tilde{\mathcal{X}}, P, \tilde{f}\rangle$ 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*

$$\tilde{\mathcal{X}} = \mathcal{X}_J \times \mathcal{X}_{J'} \times \mathcal{X}_V \times \mathcal{X}_{V'} \times \mathcal{X}_W$$

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

$$\tilde{f}_u\big((x_J, x_{J'}), (x_V, x_{V'}), x_W\big) = \begin{cases} f_u(x_J, x_V, x_W) & u \in V, \\ f_{u^\circ}(x_{J'}, x_{V'}, x_W) & u \in V'. \end{cases}$$

The twinning operation 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. 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 SCM and then intervene on it, for the second, we just intervene on the SCM and there is no need for twinning.

Hard interventions are compatible with the twinning operation, in the following sense:

Proposition 6.2.3. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be an SCM.*

- *For $T \subseteq J \cup V$, $x_T \in \mathcal{X}_T$:*
$$(\mathcal{M}^{\text{twin}})_{\text{do}(X_T = x_T, X_{T'} = x_T)} = (\mathcal{M}_{\text{do}(X_T = x_T)})^{\text{twin}}.$$
- *For $T \subseteq W$, $Q_T \in \prod_{t\in T} \mathcal{P}(\mathcal{X}_t)$:*
$$(\mathcal{M}^{\text{twin}})_{\text{do}(X_T \sim Q_T)} = (\mathcal{M}_{\text{do}(X_T \sim Q_T)})^{\text{twin}}.$$
- *For $T \subseteq J \cup V$:*
$$(\mathcal{M}^{\text{twin}})_{\text{do}(T, T')} = (\mathcal{M}_{\text{do}(T)})^{\text{twin}}.$$

Proof. These properties all follow by writing out the definitions and checking commutativity of the operations performed on the various components of the SCM tuple one-by-one. □

Another important property of the twinning operation is that it preserves unique solvability and simplicity.

Lemma 6.2.4. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be an SCM. Let $T_1 \subseteq J \cup V$, $T_2 \subseteq J' \cup V'$ and $T_3 \subseteq W$. If $g_{\text{do}(T_1 \cup T_3)} : \mathcal{X}_{J \cup T_1 \cup T_3} \times \mathcal{X}_{W \setminus T_3} \to \mathcal{X}_{V \setminus T_1}$ is a solution function of $\mathcal{M}_{\text{do}(T_1 \cup T_3)}$ and $g_{\text{do}(T_2^\circ \cup T_3)} : \mathcal{X}_{J \cup T_2^\circ \cup T_3} \times \mathcal{X}_{W \setminus T_3} \to \mathcal{X}_{V \setminus T_2^\circ}$ is a solution function of $\mathcal{M}_{\text{do}(T_2^\circ \cup T_3)}$ then*

$$\begin{aligned} \tilde{g} : \mathcal{X}_{(J \cup T_1 \cup T_3) \cup (J' \cup T_2)} \times \mathcal{X}_{W \setminus T_3} &\to \mathcal{X}_{V \setminus T_1} \times \mathcal{X}_{V' \setminus T_2} \\ : (x_{(J \cup T_1 \cup T_3) \cup (J' \cup T_2)}, x_{W \setminus T_3}) &\mapsto \big(g_{\text{do}(T_1 \cup T_3)}(x_{J \cup T_1 \cup T_3}, x_{W \setminus T_3}), g_{\text{do}(T_2^\circ \cup T_3)}(x_{J' \cup T_2 \cup T_3}, x_{W \setminus T_3})\big) \end{aligned}$$

is a solution function of $(\mathcal{M}^{\text{twin}})_{\text{do}(T_1 \cup T_2 \cup T_3)}$. In case $g_{\text{do}(T_1 \cup T_3)}$ and $g_{\text{do}(T_2^\circ \cup T_3)}$ are unique, $\tilde{g}$ is also the unique solution function of $(\mathcal{M}^{\text{twin}})_{\text{do}(T_1 \cup T_2 \cup T_3)}$.

Proof. Let $\tilde{f}$ denote the causal mechanism of $\mathcal{M}^{\text{twin}}$. For all $x \in \mathcal{X}_J \times \mathcal{X}_{J'} \times \mathcal{X}_V \times \mathcal{X}_{V'} \times \mathcal{X}_W$,

$$
\begin{aligned}
& x_{(V \cup V') \setminus (T_1 \cup T_2)} = \tilde{f}_{(V \cup V') \setminus (T_1 \cup T_2)}(x) \\
\iff & \begin{cases} x_{V \setminus T_1} & = f_{V \setminus T_1}(x_J, x_V, x_W) \\ x_{V' \setminus T_2} & = f_{(V' \setminus T_2)^\circ}(x_{J'}, x_{V'}, x_W) \end{cases} \\
\impliedby & \begin{cases} x_{V \setminus T_1} & = g_{\mathrm{do}(T_1 \cup T_3)}(x_{J \cup T_1 \cup T_3}, x_{W \setminus T_3}) \\ x_{V' \setminus T_2} & = g_{\mathrm{do}(T_2^\circ \cup T_3)}(x_{J' \cup T_2 \cup T_3}, x_{W \setminus T_3}) \end{cases} \\
\iff & x_{(V \cup V') \setminus (T_1 \cup T_2)} = \tilde{g}(x_{J \cup T_1 \cup T_3}, x_{J' \cup T_2}, x_{W \setminus T_3})
\end{aligned}
$$

In case $g_{\mathrm{do}(T_1 \cup T_3)}$ and $g_{\mathrm{do}(T_2^\circ \cup T_3)}$ are unique, the " $\impliedby$ " becomes an " $\iff$ ". □

Proposition 6.2.5. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ be an SCM. If $\mathcal{M}$ is uniquely solvable, then $\mathcal{M}^{\text{twin}}$ is uniquely solvable. Furthermore, if $\mathcal{M}$ is simple, then $\mathcal{M}^{\text{twin}}$ is simple.*

Proof. The first statement follows directly from Lemma 6.2.4 by taking $T_1 = T_2 = \emptyset$. The second statement follows from Lemma 6.2.4 in combination with Theorem 6.1.2. □

6.3. Equivalences

Equivalence relations are ubiquitous in mathematics. They capture the notion that mathematical objects can be "equivalent" from some point of view.

Definition 6.3.1. *Let Z be a set and $R \subseteq Z^2$ be a relation on Z (i.e., a subset of ordered pairs of Z). R is called an* equivalence relation *if*

1. *R is reflexive: $(a, a) \in R$ for all $a \in Z$;*
2. *R is symmetric: $(a, b) \in R \iff (b, a) \in R$ for all $a, b \in Z$;*
3. *R is transitive: if $(a, b) \in R$ and $(b, c) \in R$ then $(a, c) \in R$ for all $a, b, c \in Z$.*

In this section, we will discuss several important equivalence relations between SCMs. We define the following equivalences amongst SCMs.

Definition 6.3.2. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ and $\tilde{\mathcal{M}} = \langle J, V, W, \mathcal{X}, P, \tilde{f} \rangle$ be two SCMs that may differ only in terms of their causal mechanism. We say that $\mathcal{M}$ is* equivalent *to $\tilde{\mathcal{M}}$ and write $\mathcal{M} \equiv \tilde{\mathcal{M}}$ if for each $v \in V$,*

$$\forall x \in \mathcal{X} : \qquad x_v = f_v(x_J, x_V, x_W) \iff x_v = \tilde{f}_v(x_J, x_V, x_W).$$

In words, $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are considered equivalent if they at most differ in terms of their causal mechanisms, yet each of their structural equations has the same solutions.

Remark 6.3.3. *Equivalent SCMs are equivalent for many purposes:*

1. *Equivalence is preserved by hard interventions.*

2. *Equivalence is preserved by twinning.*

3. *Unique solvability and simplicity are invariants of equivalence.*

4. *Equivalent SCMs have the same solution functions, solutions, outcomes, and Markov kernels.*

Exercise 6.3.4. *Prove this remark.*

Example 6.3.5. *The SCM with the single structural equation*

$$X = -X^3 + X + W$$

where X is endogenous, and W is exogenous, is equivalent to the SCM obtained by replacing the structural equation with

$$X = \sqrt[3]{W}.$$

Note that X no longer appears on the r.h.s..

This notion of equivalence is rather strong.

Example 6.3.6. *The SCM with endogenous variable $X \in \mathbb{R}$ and exogenous random variable W with domain $\mathbb{R}$ and structural equation*

$$X = WX + c$$

is not equivalent to the SCM obtained by replacing the structural equation with

$$X = \begin{cases} \xi & W = 1 \\ \frac{c}{1-W} & W \neq 1 \end{cases}$$

no matter how we choose ξ, even when $P(W = 1) = 0$. If one does not model interventions on exogenous random variables, weaker notions of equivalence can be used that replace the "for all" quantifier over values of exogenous random variables by the "for almost all" quantifier (see [BFPM21]).

We often also make use of other notions of equivalence as well. For simplicity of exposition, we provide the definitions only for simple SCMs (the general definitions are provided in [BFPM21] for SCMs without exogenous input variables).

Definition 6.3.7. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ and $\tilde{\mathcal{M}} = \langle \tilde{J}, \tilde{V}, \tilde{W}, \tilde{\mathcal{X}}, \tilde{P}, \tilde{f}\rangle$ be two simple SCMs and $O \subseteq (V \cup W) \cap (\tilde{V} \cup \tilde{W})$ a subset. We say that:*

1. *$\mathcal{M}$ and $\tilde{\mathcal{M}}$ are* observationally 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:*

$$P_{\mathcal{M}}(X_O \mid \mathrm{do}(X_J)) = P_{\tilde{\mathcal{M}}}(X_O \mid \mathrm{do}(X_{\tilde{J}})).$$

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., $P_{\mathcal{M}}(X_O \mid \mathrm{do}(X_J))$ must be essentially constant in $x_{J\setminus\tilde{J}}$ and $P_{\tilde{\mathcal{M}}}(X_O \mid \mathrm{do}(X_{\tilde{J}}))$ must be essentially constant in $x_{\tilde{J}\setminus J}$.

2. *$\mathcal{M}$ and $\tilde{\mathcal{M}}$ are* interventionally equivalent w.r.t. *O if for every subset $T \subseteq O$ the intervened SCMs $\mathcal{M}_{\mathrm{do}(T)}$ and $\tilde{\mathcal{M}}_{\mathrm{do}(T)}$ are observationally equivalent w.r.t. $O \setminus T$;*

3. *$\mathcal{M}$ and $\tilde{\mathcal{M}}$ are said to be* counterfactually equivalent w.r.t. *O if the twin SCMs $\mathcal{M}^{\mathrm{twin}}$ and $\tilde{\mathcal{M}}^{\mathrm{twin}}$ are interventionally equivalent w.r.t. $O \cup O'$, where O' is the copy of $O \cap (V \cap \tilde{V})$ in $V' \cap \tilde{V}'$.*

In case the variables in O (and $J \cup \tilde{J}$) are considered observed, whereas the variables in $(V \cup W) \setminus O$ and $(\tilde{V} \cup \tilde{W}) \setminus O$ are considered latent, we sometimes omit the "w.r.t. O" in this terminology (i.e., we call $\mathcal{M}$ and $\tilde{\mathcal{M}}$ observationally equivalent, interventionally equivalent and counterfactually equivalent, respectively).

More generally, one could define interventional and counterfactual equivalence not only with respect to an observed set of variables, but also with respect to a given set of interventions.

Lemma 6.3.8. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ be a simple SCM. Then $\mathcal{M}$ and $\mathcal{M}^{\mathrm{twin}}$ are interventionally equivalent w.r.t. $V \cup W$.*

Proof. We have to show that for any $T \subseteq V \cup W$, $\mathcal{M}_{\mathrm{do}(T)}$ and $(\mathcal{M}^{\mathrm{twin}})_{\mathrm{do}(T)}$ are observationally equivalent w.r.t. $(V \cup W) \setminus T$. Let $T \subseteq V \cup W$, and write $T_1 = T \cap V$, $T_2 = \emptyset$, $T_3 = T \cap W$. Then $T = T_1 \cup T_3$.

By Lemma 6.2.4, with $g_{\mathrm{do}(T)} : \mathcal{X}_{J \cup T} \times \mathcal{X}_{W \setminus T_3} \to \mathcal{X}_{V \setminus T_1}$ the solution function of $\mathcal{M}_{\mathrm{do}(T)}$ and $g_{\mathrm{do}(T_3)} : \mathcal{X}_{J \cup T_3} \times \mathcal{X}_{W \setminus T_3} \to \mathcal{X}_V$ the solution function of $\mathcal{M}_{\mathrm{do}(T_3)}$,

$$\begin{aligned} \tilde{g} : \mathcal{X}_{(J \cup T) \cup J'} \times \mathcal{X}_{W \setminus T_3} &\to \mathcal{X}_{V \setminus T_1} \times \mathcal{X}_{V'} \\ : (x_{(J \cup T) \cup J'}, x_{W \setminus T_3}) &\mapsto \big(g_{\mathrm{do}(T)}(x_{J \cup T}, x_{W \setminus T_3}), g_{\mathrm{do}(T_3)}(x_{J' \cup T_3}, x_{W \setminus T_3})\big) \end{aligned}$$

is the unique solution function of $(\mathcal{M}^{\mathrm{twin}})_{\mathrm{do}(T)}$. Note that $g_{\mathrm{do}(T)} \circ \mathrm{pr}_{\mathcal{X}_{J \cup T} \times \mathcal{X}_{W \setminus T_3}} = \tilde{g}_{V \setminus T}$.

Then $P_{\mathcal{M}_{\mathrm{do}(T)}}(X_{(V \cup W) \setminus T} \mid \mathrm{do}(X_{J \cup T}))$ is the push-forward $(g_{\mathrm{do}(T)}, \mathrm{id}_{\mathcal{X}_{W \setminus T}})_*(P_{W \setminus T})$ of the marginal exogenous distribution $P_{W \setminus T}$ of $\mathcal{M}_{\mathrm{do}(T)}$ (interpreted as a constant Markov kernel $\mathcal{X}_{J \cup T} \dashrightarrow \mathcal{X}_{W \setminus T}$). We obtain the marginal Markov kernel $P_{(\mathcal{M}^{\mathrm{twin}})_{\mathrm{do}(T)}}(X_{(V \cup W) \setminus T} \mid \mathrm{do}(X_{J \cup T \cup J'}))$ as the push-forward $(\tilde{g}_{V \setminus T}, \mathrm{id}_{\mathcal{X}_{W \setminus T}})_*(P_{W \setminus T})$ of the marginal exogenous distribution $P_{W \setminus T}$ of $(\mathcal{M}^{\mathrm{twin}})_{\mathrm{do}(T)}$, now interpreted as a constant Markov kernel $\mathcal{X}_{J \cup T} \times \mathcal{X}_{J'} \to \mathcal{X}_{W \setminus T}$. Since $g_{\mathrm{do}(T)} \circ \mathrm{pr}_{\mathcal{X}_{J \cup T} \times \mathcal{X}_{W \setminus T_3}} = \tilde{g}_{V \setminus T}$, we obtain the desired conclusion. □

One can show the following properties of these equivalences:

Proposition 6.3.9. *For simple SCMs $\mathcal{M}, \tilde{\mathcal{M}}$ and a subset $O \subseteq (V \cap \tilde{V}) \cup (W \cap \tilde{W})$:*

1. *If $\mathcal{M} \equiv \tilde{\mathcal{M}}$ then $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are counterfactually equivalent w.r.t. O.*

2. *If $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are counterfactually equivalent w.r.t. O then $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are interventionally equivalent w.r.t. O.*

3. *If $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are interventionally equivalent w.r.t. O then $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are observationally equivalent w.r.t. O.*

However, the reverse implications do not hold in general. This 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".

Example 6.3.10 (Observational equivalence does not imply interventional equivalence)**.** *Consider the SCM $\mathcal{M}$ with*

$$N \sim \mathcal{N}(\mu, \sigma^2)$$
$$C = \alpha + \beta N$$

and the SCM $\tilde{\mathcal{M}}$ with

$$C \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$$
$$N = \tilde{\alpha} + \tilde{\beta} C$$

These SCMs are simple and their observational distributions are respectively

$$P(N, C) = \mathcal{N}\left(\begin{pmatrix} \mu \\ \alpha\mu \end{pmatrix}, \begin{pmatrix} \sigma^2 & \beta\sigma^2 \\ \beta\sigma^2 & \beta^2\sigma^2 \end{pmatrix}\right)$$

and

$$P(N, C) = \mathcal{N}\left(\begin{pmatrix} \tilde{\alpha}\tilde{\mu} \\ \tilde{\mu} \end{pmatrix}, \begin{pmatrix} \tilde{\beta}^2\tilde{\sigma}^2 & \tilde{\beta}\tilde{\sigma}^2 \\ \tilde{\beta}\tilde{\sigma}^2 & \tilde{\sigma}^2 \end{pmatrix}\right)$$

For certain parameter choices, they are observationally equivalent. However, they are not interventionally equivalent except for very special parameter choices.

In general, interventional equivalence does not imply counterfactual equivalence. Even interventionally equivalent SCMs with the same causal mechanism (that differ only in terms of their exogenous distributions) may not be counterfactually equivalent. For example, the SCMs $\mathcal{M}_\rho$ and $\mathcal{M}_{\rho'}$ with $\rho \neq \rho'$ in the following example are interventionally equivalent, but not counterfactually equivalent.

Example 6.3.11 (Interventional equivalence does not imply counterfactual equivalence [Daw02])**.** *For parameter $\rho \in [0, 1]$, consider the SCM $\mathcal{M}_\rho$ with binary exogenous input variable $X \in \{0, 1\}$, endogenous variable Y, 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 SCM could be used to model whether a patient was treated or not (X) and the corresponding outcome for that patient (Y).

Suppose in the actual world we did not assign treatment to a patient ($X = 0$) and the outcome was $Y^{\mathrm{do}(0)} = y \in \mathbb{R}$. 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 SCM $\mathcal{M}_\rho^{\text{twin}}$. The counterfactual query then asks for $P((Y')^{\text{do}(1)} = y' \mid Y^{\text{do}(0)} = y)$, where Y^0 is the factual outcome, and $(Y')^{\text{do}(1)}$ is the counterfactual outcome (which are both marginal potential outcomes of the twinned SCM). One can calculate that

$$P\big((Y')^{\text{do}(1)}, Y^{\text{do}(0)}\big) = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$

and hence $P((Y')^{\text{do}(1)} \mid Y^{\text{do}(0)} = y) = \mathcal{N}(\rho y, 1-\rho^2)$ (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 $P(Y \mid \text{do}(X))$, since this 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, SCMs $\mathcal{M}_\rho$ and $\mathcal{M}_{\rho'}$ with $\rho \neq \rho'$ are interventionally equivalent, but not counterfactually equivalent.

The lesson of this example is that if one attempts to learn an SCM 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 queries. In other words, data-driven estimation of counterfactuals can be an ill-posed problem. Nevertheless, counterfactuals 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 is on very slippery terrain when it comes to answering such questions. Unless one is willing to make very strong modeling assumptions, giving objective answers (even probabilistic ones) can be an impossible task. In certain cases, though, it is possible to derive *bounds* on counterfactual queries from the observational and interventional Markov kernels.

6.4. Marginalizations

When modeling a system, we sometimes want to "hide" details of a subsystem. The following operation on SCMs 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 SCM 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 6.4.1. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be an SCM and $L \subseteq V$ such that $\mathcal{M}_{\mathrm{do}(V\setminus L)}$ is uniquely solvable. Write $M = V \setminus L$ and let $\tilde{g}_L : \mathcal{X}_{J\cup M} \times \mathcal{X}_W \to \mathcal{X}_L$ be the unique solution function for $\mathcal{M}_{\mathrm{do}(M)}$. Then we call $\mathcal{M}_{\setminus L} = \langle J, M, W, \mathcal{X}_J \times \mathcal{X}_M \times \mathcal{X}_W, P, \tilde{f}\rangle$ with*

$$\tilde{f}(x_J, x_M, x_W) = f_M(x_J, x_M, \tilde{g}_L(x_J, x_M, x_W), x_W)$$

the marginalization of $\mathcal{M}$ over L.

For simple SCMs, marginalizations are obviously defined over any subset $L \subseteq V$. Marginalization preserves unique solvability, as the following lemma shows.

Lemma 6.4.2. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be an SCM and $L \subseteq V$ such that $\mathcal{M}_{\mathrm{do}(V\setminus L)}$ is uniquely solvable. If $\mathcal{M}$ is uniquely solvable then its marginalization $\mathcal{M}_{\setminus L}$ is uniquely solvable, and the unique solution function for $\mathcal{M}_{\setminus L}$ is just $g_M = \mathrm{pr}_M \circ g$, where $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ is the unique solution function for $\mathcal{M}$, $M = V \setminus L$ and $\mathrm{pr}_M : \mathcal{X}_V \to \mathcal{X}_M : x \mapsto x_M$ is the canonical projection on M.*

Proof. Let $\tilde{g}_L : \mathcal{X}_{J\cup M} \times \mathcal{X}_W \to \mathcal{X}_L$ be the unique solution function for $\mathcal{M}_{\mathrm{do}(M)}$. From the properties of the solution functions, we derive that for all $x \in \mathcal{X}$,

$$\begin{cases} x_L = g_L(x_J, x_W) \\ x_M = g_M(x_J, x_W) \end{cases} \iff x_V = g(x_J, x_W) \iff x_V = f(x) \iff \begin{cases} x_L = f_L(x) \\ x_M = f_M(x) \end{cases}$$

$$\iff \begin{cases} x_L = \tilde{g}_L(x_J, x_M, x_W) \\ x_M = f_M(x_J, x_M, x_L, x_W) \end{cases} \iff \begin{cases} x_L = \tilde{g}_L(x_J, x_M, x_W) \\ x_M = f_M(x_J, x_M, \tilde{g}_L(x_J, x_M, x_W), x_W) \end{cases}$$

$$\iff \begin{cases} x_L = \tilde{g}_L(x_J, x_M, x_W) \\ x_M = \tilde{f}(x_J, x_M, x_W) \end{cases}$$

Hence for all $x_J \in \mathcal{X}_J, x_W \in \mathcal{X}_W, x_M \in \mathcal{X}_M$:

$$x_M = g_M(x_J, x_W) \iff x_M = \tilde{f}(x_J, x_M, x_W).$$

Therefore, $g_M = \mathrm{pr}_M \circ g$ is the unique solution function for $\mathcal{M}_{\setminus L}$, and the marginalized SCM $\mathcal{M}_{\setminus L}$ is uniquely solvable. □

Remark 6.4.3. *This also directly implies that if $\mathcal{M}$ is uniquely solvable and its marginalization $\mathcal{M}_{\setminus L}$ over $L \subseteq V$ is defined, the observational Markov kernel of the marginalization is obtained by marginalizing the original observational Markov kernel:*

$$P_{\mathcal{M}_{\setminus L}}(X_{V\setminus L}, X_W \mid \mathrm{do}(X_J)) = P_{\mathcal{M}}(X_{V\setminus L}, X_W \mid \mathrm{do}(X_J)).$$

Under certain conditions, hard interventions and marginalization commute (i.e., it does not matter in which order we apply them).

Proposition 6.4.4. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be an SCM. For $L \subseteq V$ such that $\mathcal{M}_{\mathrm{do}(V\setminus L)}$ is uniquely solvable, and a hard intervention $\mathrm{do}(T \dots)$ with target $T \subseteq J \cup W \cup V$ (of any of the three variants) such that $L \cap T = \emptyset$:*

$$(\mathcal{M}_{\mathrm{do}(T\dots)})_{\setminus L} = (\mathcal{M}_{\setminus L})_{\mathrm{do}(T\dots)}.$$

Proof. This follows by writing out the definitions and checking commutativity of the operations performed on the various components of the SCM tuple one-by-one. □

Marginalization is also compatible with the twinning operation.

Proposition 6.4.5. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be an SCM. For $L \subseteq V$ such that $\mathcal{M}_{\mathrm{do}(V\setminus L)}$ is uniquely solvable,*

$$(\mathcal{M}_{\setminus L})^{\mathrm{twin}} = (\mathcal{M}^{\mathrm{twin}})_{\setminus (L\cup L')}.$$

Proof. This follows by writing out the definitions and checking commutativity of the operations performed on the various components of the SCM tuple one-by-one. □

We will show that for simple SCMs, 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 6.4.6. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be an SCM and $L_1, L_2 \subseteq V$ such that $L_1 \cap L_2 = \emptyset$. If $\mathcal{M}_{\mathrm{do}(V\setminus L_1)}$ is uniquely solvable, and $(\mathcal{M}_{\setminus L_1})_{\mathrm{do}((V\setminus L_1)\setminus L_2)}$ is uniquely solvable, then $\mathcal{M}_{\mathrm{do}(V\setminus(L_1\cup L_2))}$ is uniquely solvable, 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:*

$$(\mathcal{M}_{\setminus L_1})_{\setminus L_2} = \mathcal{M}_{\setminus (L_1\cup L_2)}.$$

Proof. Write $M_1 = V \setminus L_1$. Let $\tilde{g} : \mathcal{X}_{J\cup M_1} \times \mathcal{X}_W \to \mathcal{X}_{L_1}$ be the unique solution function of $\mathcal{M}_{\mathrm{do}(M_1)}$, i.e., for all $x \in \mathcal{X}$:

$$x_{L_1} = \tilde{g}(x_J, x_{M_1}, x_W) \iff x_{L_1} = f_{L_1}(x).$$

Let $\tilde{f} : \mathcal{X}_J \times \mathcal{X}_{M_1} \times \mathcal{X}_W \to \mathcal{X}_{M_1}$ with

$$\tilde{f}(x_J, x_{M_1}, x_W) = f_{M_1}(x_J, x_{M_1}, \tilde{g}(x_J, x_{M_1}, x_W), x_W)$$

be the causal mechanism of the marginal SCM $\mathcal{M}_{\setminus L_1}$.

If $(\mathcal{M}_{\setminus L_1})_{\mathrm{do}(V\setminus(L_1\cup L_2))}$ is uniquely solvable, it has a unique solution function $\tilde{\tilde{g}} : \mathcal{X}_J \times \mathcal{X}_{V\setminus(L_1\cup L_2)} \times \mathcal{X}_W \to \mathcal{X}_{L_2}$, i.e., for all $x \in \mathcal{X}$:

$$x_{L_2} = \tilde{\tilde{g}}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) \iff x_{L_2} = \tilde{f}_{L_2}(x_J, x_{L_2}, x_{M_1\setminus L_2}, x_W)$$

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

$$\begin{aligned} h_{L_1}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) &= \tilde{g}(x_J, x_{M_1\setminus L_2}, \tilde{\tilde{g}}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W), x_W) \\ h_{L_2}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) &= \tilde{\tilde{g}}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) \end{aligned}$$

Then for all $x \in \mathcal{X}$:

$$\begin{aligned} & \begin{cases} x_{L_1} &= f_{L_1}(x) \\ x_{L_2} &= f_{L_2}(x) \end{cases} \\ \iff & \begin{cases} x_{L_1} &= \tilde{g}(x_J, x_{M_1}, x_W) \\ x_{L_2} &= f_{L_2}(x_J, x_{M_1}, x_{L_1}, x_W) \end{cases} \\ \iff & \begin{cases} x_{L_1} &= \tilde{g}(x_J, x_{M_1}, x_W) \\ x_{L_2} &= f_{L_2}(x_J, x_{M_1}, \tilde{g}(x_J, x_{M_1}, x_W), x_W) \end{cases} \\ \iff & \begin{cases} x_{L_1} &= \tilde{g}(x_J, x_{M_1\setminus L_2}, x_{L_2}, x_W) \\ x_{L_2} &= \tilde{\tilde{g}}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) \end{cases} \\ \iff & \begin{cases} x_{L_1} &= \tilde{g}(x_J, x_{M_1\setminus L_2}, \tilde{\tilde{g}}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W), x_W) \\ x_{L_2} &= \tilde{\tilde{g}}(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) \end{cases} \\ \iff & x_{L_1\cup L_2} = h(x_J, x_{V\setminus(L_1\cup L_2)}, x_W) \end{aligned}$$

where in the first equivalence we used the unique solvability of $\mathcal{M}_{\mathrm{do}(M_1)}$, in the second equivalence we used substitution, in the third equivalence we used the unique solvability of $(\mathcal{M}_{\setminus L_1})_{\mathrm{do}(V\setminus(L_1\cup L_2))}$, in the fourth equivalence we used substitution again, and in the fifth equivalence we used the definition of h. Therefore, h is the unique solution function for $\mathcal{M}_{\mathrm{do}(V\setminus(L_1\cup L_2))}$, which must therefore be uniquely solvable. By checking the definition, one concludes that $(\mathcal{M}_{\setminus L_1})_{\setminus L_2} = \mathcal{M}_{\setminus(L_1\cup L_2)}$. □

Proposition 6.4.7. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM. For any $L \subseteq V$, its marginalization $\mathcal{M}_{\setminus L}$ is also simple.*

Proof. Let $T \subseteq J \cup W \cup V \setminus L$. By Proposition 6.4.4, the marginalization commutes with the hard intervention:

$$(\mathcal{M}_{\setminus L})_{\mathrm{do}(T)} = (\mathcal{M}_{\mathrm{do}(T)})_{\setminus L}.$$

Because $\mathcal{M}$ is simple, also $\mathcal{M}_{\mathrm{do}(T)}$ is simple (by Proposition 5.7.7), and in particular it is uniquely solvable. From Lemma 6.4.2 it follows that also its marginalization $(\mathcal{M}_{\mathrm{do}(T)})_{\setminus L}$ is uniquely solvable. This means that $(\mathcal{M}_{\setminus L})_{\mathrm{do}(T)}$ is uniquely solvable. Since this holds for any $T \subseteq J \cup W \cup V \setminus L$, $\mathcal{M}_{\setminus L}$ is simple. □

Corollary 6.4.8. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM. For $L_1, L_2 \subseteq V$ with $L_1 \cap L_2 = \emptyset$,*

$$(\mathcal{M}_{\setminus L_1})_{\setminus L_2} = (\mathcal{M}_{\setminus L_2})_{\setminus L_1} = \mathcal{M}_{\setminus (L_1 \cup L_2)}.$$

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 [BFPM21], we will here only prove this for simple SCMs.

Theorem 6.4.9. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM, $L \subseteq V$, and $\mathcal{M}_{\setminus L}$ its marginalization over L. Then $\mathcal{M}$ and $\mathcal{M}_{\setminus L}$ are observationally, interventionally and counterfactually equivalent w.r.t. $(V \cup W) \setminus L$.*

Proof. Write $M = V \setminus L$. We first show that the marginal Markov kernels $P_{\mathcal{M}}(X_M, X_W \mid \mathrm{do}(X_J))$ and $P_{\mathcal{M}_{\setminus L}}(X_M, X_W \mid \mathrm{do}(X_J))$ are the same. The former is obtained as:

$$P_{\mathcal{M}}(X_M, X_W \mid \mathrm{do}(X_J)) = (\mathrm{pr}_{M\cup W} \circ (g, \mathrm{id}_{\mathcal{X}_W}))_*(P) = (g_M, \mathrm{id}_{\mathcal{X}_W})_*(P),$$

where $g : \mathcal{X}_J \times \mathcal{X}_W \to \mathcal{X}_V$ is the unique solution function of $\mathcal{M}$ and $\mathrm{pr}_{M\cup W} : \mathcal{X}_{V\cup W} \to \mathcal{X}_{M\cup W}$ is the canonical projection on the $M \cup W$ components. The latter is obtained as:

$$P_{\mathcal{M}_{\setminus L}}(X_M, X_W \mid \mathrm{do}(X_J)) = (g_M, \mathrm{id}_{\mathcal{X}_W})_*(P)$$

since by Lemma 6.4.2, g_M is the (unique) solution function of $\mathcal{M}_{\setminus L}$. This means that both push-forwards are identical.

Let $T \subseteq M \cup W$. Then $(\mathcal{M}_{\setminus L})_{\mathrm{do}(T)} = (\mathcal{M}_{\mathrm{do}(T)})_{\setminus L}$ by Proposition 6.4.4. The observational equivalence of $\mathcal{M}_{\mathrm{do}(T)}$ and $(\mathcal{M}_{\mathrm{do}(T)})_{\setminus L}$ w.r.t. $(M \cup W) \setminus T$ hence implies the observational equivalence of $\mathcal{M}_{\mathrm{do}(T)}$ and $(\mathcal{M}_{\setminus L})_{\mathrm{do}(T)}$ w.r.t. $(M \cup W) \setminus T$. Since this holds for all $T \subseteq M \cup W$, $\mathcal{M}$ and $\mathcal{M}_{\setminus L}$ are interventionally equivalent w.r.t. $M \cup W$.

Finally, $(\mathcal{M}_{\setminus L})^{\mathrm{twin}} = (\mathcal{M}^{\mathrm{twin}})_{\setminus (L \cup L')}$ by Proposition 6.4.5. Since $\mathcal{M}^{\mathrm{twin}}$ and its marginalization $(\mathcal{M}^{\mathrm{twin}})_{\setminus (L\cup L')}$ are interventionally equivalent w.r.t. $M \cup M' \cup W$, $\mathcal{M}$ and $\mathcal{M}_{\setminus L}$ are counterfactually equivalent w.r.t. $M \cup W$. □

So the marginalization operation indeed effectively hides the details of a subsystem, while preserving all causal semantics on the remaining part.

6.5. Graphs of SCMs

A useful abstraction of an SCM is its graph. The directed edges in the graph of an SCM will express the following "parent"-relation that captures *functional dependencies* in the structural equations / causal mechanisms.

Definition 6.5.1. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be an SCM. For $i \in J \cup V \cup W$ and $j \in V$, we say that i* is a parent of j according to $\mathcal{M}$ *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 for all $x \in \mathcal{X}$,*

$$x_j = f_j(x) \iff x_j = \tilde{f}_j(x_{\setminus i}),$$

where $x_{\setminus i}$ is shorthand for $x_{(J\cup V\cup W)\setminus\{i\}}$.

In words, i is parent of j if the causal mechanism for j is equivalent to one that is constant with respect to its input component i. By definition, exogenous (input and random) variables have no parents. Note that the parent-relationship is preserved under equivalence. Using directed edges to encode the parent-relationship, we define the graph of the SCM.

Definition 6.5.2. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ be an SCM. The CDG $\langle J, V \cup W, E \rangle$ with input nodes J, output nodes $V \cup W$, and directed edges*

$$E = \{i \to j : i \in J \cup W \cup V, j \in V : i \text{ is parent of } j \text{ according to } \mathcal{M}\}$$

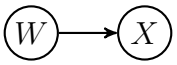
is called the graph of the SCM *and will be denoted as $G(\mathcal{M})$.*

Since the parent-relationship is preserved under equivalence, equivalent SCMs have the same graphs. However, observationally equivalent SCMs may have different graphs, and even interventionally equivalent SCMs may have different graphs.

Example 6.5.3. *The SCM in Example 6.3.5, with endogenous variable X and exogenous random variable W and structural equation*

$$X = -X^3 + X + W$$

has graph:



It does not have a self-cycle at X, since X is not a parent to itself: the structural equation is equivalent to

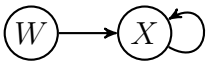
$$X = \sqrt[3]{W},$$

where X does not appear on the r.h.s..

On the other hand, the graph of the SCM in Example 6.3.6 with structural equation

$$X = WX + c$$

does have a self-cycle at X:



Self-cycles are a warning sign of complications with respect to solvability. They are of no concern when restricting to the class of simple SCMs.

Proposition 6.5.4. *For $j \in V$, we have that there is a self-cycle $j \to j$ in $G(\mathcal{M})$ if and only if $\mathcal{M}_{\text{do}(V \setminus \{j\})}$ is* not *uniquely solvable. In particular, graphs of simple SCMs have no self-cycles.*

While the graph explicitly represents all exogenous random variables, coarser representations obtained via (graphical) marginalization are also useful.

Remark 6.5.5. *An often used convention is to consider* all *exogenous random variables to be latent. In that case, a useful graph to consider is the marginalization* $(G)^{\backslash W}(\mathcal{M})$ *of the graph* $G(\mathcal{M})$*. The marginal CDMG* $G^{\backslash W}(\mathcal{M}) = \langle J, V, E, L\rangle$ *has input nodes* J*, output nodes* V*, directed edges*

$$E = \{i \to j : i \in J \cup V, j \in V : i \text{ is parent of } j \text{ according to } \mathcal{M}\}$$

and bidirected edges

$$L = \{j \leftrightarrow k : j \in V, k \in V, j \neq k : j \text{ and } k \text{ have a common parent } i \in W \text{ according to } \mathcal{M}\}.$$

In other words, $G^{\backslash W}(\mathcal{M})$ *is obtained from* $G(\mathcal{M})$ *by replacing the nodes representing the exogenous random variables and their outgoing directed edges with bidirected edges, i.e., any pattern* $\leftarrow i \rightarrow$ *with* $i \in W$ *is replaced by* $\leftrightarrow$*.*

We already provided definitions for hard interventions on graphs (Definition 3.2.1) and for marginalizations (latent projections) of graphs (Definition 3.2.1). We can also define a twinning operation on graphs. We will only define this for graphs without bidirected edges.

Definition 6.5.6. *Let* $G = \langle J, V \cup W, E\rangle$ *be a CDG such that* $\mathrm{Pa}^G(W) = \emptyset$*. Write* $J' := \{j' : j \in J\}$ *and* $V' := \{v' : v \in V\}$ *for copies of* J *and* V*, respectively. The* twinned graph $G^{\mathrm{twin}(J,V)}$ *is defined as the CDG* $\langle J \dot\cup J', V \dot\cup V' \cup W, \tilde{E}\rangle$ *with directed edges*

$$\tilde{E} := E \cup \{w \to v' : w \in W, v \in V, w \to v \in E\} \cup \{i' \to v' : i \in J \cup V, v \in V, i \to v \in E\}.$$

In words, we copy the nodes $J \cup V$ *(but not the nodes* W*) and copy the edges accordingly.*

The mapping that maps an SCM to its graph is compatible with the elementary operations on SCMs.

Proposition 6.5.7. *Let* $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ *be an SCM. Then*

- *Hard interventions: for* $T \subseteq J \cup V \cup W$*,*

$$G(\mathcal{M}_{\mathrm{do}(T)}) = G(\mathcal{M})_{\mathrm{do}(T)}.$$

- *Twinning:*

$$G(\mathcal{M})^{\mathrm{twin}(J\cup V)} = G(\mathcal{M}^{\mathrm{twin}}).$$

- *Marginalizations: If* $\mathcal{M}$ *is simple, then for* $L \subseteq V$*,*

$$G(\mathcal{M}_{\backslash L}) \subseteq G(\mathcal{M})^{\backslash L}.$$

Proof. The first two statements follow by writing out the definitions. The third one is somewhat more involved. We will first prove it in case $L = \{\ell\}$ consists of a single node.

Let $G = G(\mathcal{M})$. 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 for all $x \in \mathcal{X}$:

$$x_\ell = f_\ell(x) \iff x_\ell = \tilde{f}_\ell(x_{\mathrm{Pa}^G(\ell)}).$$

Since $\ell \notin \mathrm{Pa}^G(\ell)$ because $\mathcal{M}$ is simple, the unique solution function $\tilde{g}_\ell : \mathcal{X}_{J \cup V \setminus \{\ell\} \cup W} \to \mathcal{X}_\ell$ of $\mathcal{M}_{\mathrm{do}(V \setminus \{\ell\})}$ satisfies $\tilde{g}_\ell(x_{\setminus \ell}) = \tilde{f}_\ell(x_{\mathrm{Pa}^G(\ell)})$, i.e., it only depends on the parents of ℓ. When constructing the marginalized causal mechanism for $\mathcal{M}_{\setminus\{\ell\}}$, we substitute $\tilde{g}_\ell(x_{\setminus \ell})$ into the ℓ'th input of the causal mechanism f_j of $\mathcal{M}$, 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 $\tilde{G} = G(\mathcal{M}_{\setminus \ell})$. 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 $G(\mathcal{M}_{\setminus\{\ell\}}) \subseteq G(\mathcal{M})^{\setminus\{\ell\}}$. For the general case, we can make use of induction and the facts that both for graphs and simple SCMs, we can obtain a marginalization over a subset by repeatedly marginalizing out a single remaining node in the subset, in arbitrary order. □

A subclass of SCMs that is often considered are acyclic SCMs.

Definition 6.5.8. *An SCM $\mathcal{M}$ is called* acyclic *if its graph $G(\mathcal{M})$ is acyclic.*

If one models static systems, then using acyclic SCMs rules out the presence of causal cycles (e.g., feedback loops) in the system. Acyclic SCMs are a subclass of the more general class of simple SCMs.

Proposition 6.5.9. *Acyclic SCMs are simple.*

Proof. We first show that acyclic SCMs are uniquely solvable. Let $\mathcal{M}$ be an acyclic SCM. Its graph $G := G(\mathcal{M})$ 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. Since f_v can be rewritten to be constant in the non-parents of v (similar to how this was done in the proof of Proposition 6.5.7), we can 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 uniqueness of $g : \mathcal{X}_{J \cup W} \to X_V$.

Next consider $\mathcal{M}_{\mathrm{do}(T)}$, the intervened SCM for a hard intervention on $\mathcal{M}$ with target $T \subseteq V \cup W \cup J$. It has graph $G(\mathcal{M}_{\mathrm{do}(T)}) = G_{\mathrm{do}(T)}$, 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 $\mathcal{M}_{\mathrm{do}(T)}$ is uniquely solvable. Since this holds for all targets $T \subseteq V \cup W \cup J$, we conclude that $\mathcal{M}$ is simple. □

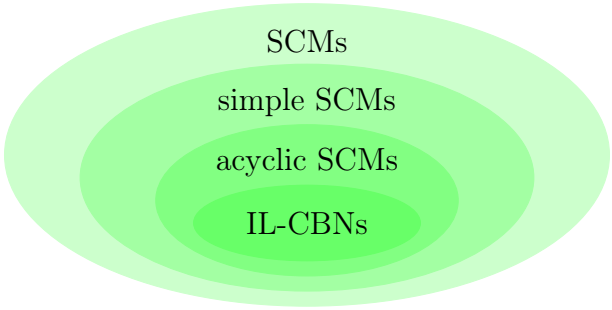


Figure 5: Venn diagram for different causal modeling classes.

Figure 5 shows a Venn diagram to illustrate how the different classes of causal models that we introduced are related. Acyclic SCMs are more expressive than IL-CBNs, because they also model counterfactuals. Simple SCMs are more expressive than acyclic SCMs because they can model (sufficiently weak) causal cycles. SCMs 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 SCMs form a "sweet spot" in the sense that they allow cyclic relationships yet their theory is not much more complicated than that of acyclic SCMs: the main difference consists in replacing d-separation with σ-separation.

6.6. (Alternative definition of σ-blocking)

PF: Alernative definition of σ-blockking copy-pasted to Graph Theory in the beginning. Here commented out.

6.7. Acyclifications

PF: Graph part of acyclification copy-pasted to Graph Theory in the beginning. Here commented out.

We can also define an operation with the same name on SCMs.

Definition 6.7.1 (Acyclification of SCM)**.** *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM with graph $G = G(\mathcal{M})$. For each strongly connected component C of $G(\mathcal{M})$ (i.e., a set of the form $\mathrm{Sc}^G(v)$ for $v \in V$), let $g_C : \mathcal{X}_{J\cup(V\setminus C)\cup W} \to \mathcal{X}_C$ be the unique solution function for $\mathcal{M}_{\mathrm{do}(V\setminus C)}$. Define $\tilde{f} : \mathcal{X}_J \times \mathcal{X}_V \times \mathcal{X}_W \to \mathcal{X}_V$ by its components*

$$\tilde{f}_v(x_J, x_V, x_W) = (g_{\mathrm{Sc}^G(v)})_v(x_J, x_{V\setminus C}, x_W)$$

for $v \in V$. $\mathcal{M}^{\mathrm{acy}} = \langle J, V, W, \mathcal{X}, P, \tilde{f}\rangle$ is called the acyclification of $\mathcal{M}$.

The crucial property of this definition is the following result, which also motivates its name.

Proposition 6.7.2. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ be a simple SCM. Its acyclification $\mathcal{M}^{\text{acy}}$ is acyclic and observationally equivalent to $\mathcal{M}$.*

Proof. We construct a directed graph S from G with its strongly connected components $\{\text{Sc}^G(v) : v \in V \cup J \cup W\}$ as nodes, and directed edges $C \to D$ if there is a directed edge $c \to d$ 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 $\text{Pred}^S_<(C)$. This implies that for $v \in V$, $C = \text{Sc}^G(v)$ can only have incoming edges in G from $\bigcup \text{Pred}^S_<(C)$. That implies that the causal mechanism f_C can only depend on variables in $\bigcup \text{Pred}^S_<(C)$, and hence the unique solution function g_C, and therefore $\tilde{f}_C$, can depend on variables in $\bigcup \text{Pred}^S_<(C)$ only. Therefore, for $v, w \in V$, a directed edge $w \to v$ in $G(\mathcal{M}^{\text{acy}})$ implies $w \in \bigcup \text{Pred}^S_<(\text{Sc}^G(v))$. We can therefore refine the topological ordering $<$ of S to a topological ordering of $G(\mathcal{M}^{\text{acy}})$, by arbitrarily ordering the nodes within each strongly connected component of G. Hence $G(\mathcal{M}^{\text{acy}})$ is acyclic.

$\mathcal{M}$ and $\mathcal{M}^{\text{acy}}$ are observationally equivalent by construction: for all $x \in \mathcal{X}$,

$$\begin{aligned}
&x = \tilde{f}(x) \\
&\iff \forall C \in S \cap V : x_C = \tilde{f}_C(x) \\
&\iff \forall C \in S \cap V : x_C = g_C(x_J, x_{V \setminus C}, x_W) \\
&\iff \forall C \in S \cap V : x_C = f_C(x) \\
&\iff x = f(x).
\end{aligned}$$

□

The notion of acyclification of an SCM is compatible with that of a graph:

Proposition 6.7.3. *Let $\mathcal{M}$ be a simple SCM. Then $G(\mathcal{M}^{\text{acy}}) \subseteq G'$ for any acyclification G' of $G(\mathcal{M})$.*

Proof. By definition, the two graphs have the same nodes (input nodes J and output nodes $V \cup W$). $G(\mathcal{M}^{\text{acy}})$ has no bidirected edges, but G' might. If there is a directed edge $i \to j$ in $G(\mathcal{M}^{\text{acy}})$ with $i \in J \cup V \cup W$ and $j \in V$, then the solution function $g_{\text{Sc}^G(j)}$ of $\mathcal{M}_{\text{do}(V \setminus \text{Sc}^G(j))}$ 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 $\mathcal{M}$ with $k \in \text{Sc}^G(j)$, i.e., if $i \to k$ in $G(\mathcal{M})$. In that case, $i \to j$ in G' by definition of the graphical acyclification. □

Hence, two nodes in the same strongly connected component of $G(\mathcal{M})$ do not have any edge between them in $G(\mathcal{M}^{\text{acy}})$, whereas they necessarily have a connecting edge in any acyclification G' of $G(\mathcal{M})$. For two nodes in different strongly connected components of $G(\mathcal{M})$, the edges in $G(\mathcal{M}^{\text{acy}})$ are also present in G', but not necessarily vice versa, as some parent-relations may cancel out in the solution function.

6.8. Global Markov property for simple SCMs

With the help of the acyclifications, we can easily derive a Markov property for simple SCMs from the Markov property for CBNs by reducing the general cyclic case to an acyclic case.

Corollary 6.8.1. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM with graph $G(\mathcal{M})$ and observational Markov kernel $P_{\mathcal{M}}(X_V, X_W \,|\, \mathrm{do}(X_J))$. Then for all $A, B, C \subseteq J \cup V \cup W$ (not necessarily disjoint) we have the implication:*

$$A \overset{\sigma}{\underset{G(\mathcal{M})}{\perp}} B \,|\, C \implies X_A \underset{P_{\mathcal{M}}(X_V, X_W \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C.$$

If one wants to make the implicit dependence on J more explicit one can equivalently also write:

$$A \overset{\sigma}{\underset{G(\mathcal{M})}{\perp}} J \cup B \,|\, C \implies X_A \underset{P_{\mathcal{M}}(X_V, X_W \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_J, X_B \,|\, X_C.$$

Proof. Choose an acyclification G' of $G(\mathcal{M})$. Then:

$$\begin{aligned}
A \overset{\sigma}{\underset{G(\mathcal{M})}{\perp}} B \,|\, C &\iff A \overset{d}{\underset{G'}{\perp}} B \,|\, C \\
&\implies A \overset{d}{\underset{G(\mathcal{M}^{\mathrm{acy}})}{\perp}} B \,|\, C \\
&\implies X_A \underset{P_{\mathcal{M}^{\mathrm{acy}}}(X_V, X_W \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C \\
&\iff X_A \underset{P_{\mathcal{M}}(X_V, X_W \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C.
\end{aligned}$$

For the various implications / equivalences, we used:

1. G' is an acyclification of $G(\mathcal{M})$ together with Proposition 3.4.2;
2. $G(\mathcal{M}^{\mathrm{acy}}) \subseteq G'$ from Proposition 6.7.3, and that removing edges cannot turn a d-separation into a d-connection;
3. the global Markov property Theorem 4.2.1 for $\mathcal{M}^{\mathrm{acy}}$ interpreted as a causal Bayesian network (with deterministic Markov kernels for the endogenous variables, and purely probabilistic Markov kernels for the exogenous random variables), exploiting Proposition 6.7.2 that states that the acyclification $\mathcal{M}^{\mathrm{acy}}$ is acyclic,
4. by Proposition 6.7.2, the acyclification $\mathcal{M}^{\mathrm{acy}}$ has the same observational Markov kernel as the original SCM $\mathcal{M}$.

□

This Markov property is not complete: it does not exploit all information in $G(\mathcal{M})$ that it possibly could to deduce conditional independences. For example, endogenous variables without parents (for which the corresponding outcomes must be constant) could additionally be taken to block walks.

6.9. Do-calculus for simple SCMs

With the global Markov property for simple SCMs, it becomes straightforward to derive the do-calculus for simple SCMs. First we will introduce some notation. The setting will be that a simple SCM $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ is given. As an auxiliary way to model hard interventions $\mathrm{do}(b)$ for $b \in B \subseteq V \cup W$ that leads to easy rules in the do-calculus, we introduce additional intervention variables I_b for $b \in B$. We denote $I_B := (I_b)_{b\in B}$. This gives an extended SCM $\tilde{\mathcal{M}} = \langle J \cup I_B, V, W, \mathcal{X} \times \prod_{b\in B}(\mathcal{X}_b \dot\cup \{\star\}), P, \tilde f\rangle$ with causal mechanism with components

$$\tilde f_b(x_{V\cup J\cup W}, x_{I_B}) := \begin{cases} f_b(x_{V\cup J\cup W}) & x_{I_b} = \star \\ x_{I_b} & x_{I_b} \in \mathcal{X}_b \end{cases}$$

for $b \in B$, and $\tilde f_v(x) := f_v(x_{V\cup J\cup W})$ for $v \in V \setminus B$. Here, $x_{I_b} = \star$ 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. We will denote the graph of $\mathcal{M}$ by G and the graph of $\tilde{\mathcal{M}}$ by $G_{\mathrm{do}(I_B)}$. In the do-calculus, we make use of the extended graph $G_{\mathrm{do}(I_B)}$ to test the separation statement, while the conclusion about the existence and identity of particular Markov kernels concerns those of the original SCM $\mathcal{M}$.

Remember the notation of Markov kernels for simple SCMs from Definition 6.1.6. Its observational Markov kernel is denoted $P_{\mathcal{M}}(X_{V\cup W} \mid \mathrm{do}(X_J))$. For a subset $T \subseteq V \cup W$, we write

$$P_{\mathcal{M}}(X_{(V\cup W)\setminus T} \mid \mathrm{do}(X_J, X_T)) := P_{\mathcal{M}_{\mathrm{do}(T)}}(X_{(V\cup W)\setminus T} \mid \mathrm{do}(X_J, X_T)).$$

By conditioning on a subset $S \subseteq (V \cup W) \setminus T$, we obtain the conditional Markov kernel

$$P_{\mathcal{M}}(X_{(V\cup W)\setminus(T\cup S)} \mid \mathrm{do}(X_J, X_T), X_S).$$

The only modification to the do-calculus for simple SCMs as compared to that for causal Bayesian networks is that we have to replace all d-separations by σ-separations.

Corollary 6.9.1 (Do-calculus for simple SCMs)**.** *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM with graph $G = G(\mathcal{M})$. Let $A, B, C \subseteq V \cup W$ and $D \subseteq V \cup W \cup J$ be such that A, B, C, D are pairwise disjoint. Then we have the following 3 rules relating marginal conditional interventional Markov kernels of $\mathcal{M}$:*

1. Insertion/deletion of observation*: If we have:*

$$A \overset{\sigma}{\underset{G_{\mathrm{do}(D)}}{\perp}} B \,|\, C \cup D,$$

then there exists a Markov kernel:

$$P\left(X_A | \cancel{X_B}, X_C, \mathrm{do}(X_{D\cup J})\right) : \mathcal{X}_C \times \mathcal{X}_D \dashrightarrow \mathcal{X}_A$$

that is a version of:

$$P_{\mathcal{M}}\left(X_A | X_{B_2}, X_C, \mathrm{do}(X_{D\cup J})\right),$$

for every subset $B_2 \subseteq B$ simultaneously. In short we could write:

$$P_{\mathcal{M}}\left(X_A | X_B, X_C, \operatorname{do}(X_{D\cup J})\right) = P_{\mathcal{M}}\left(X_A | X_C, \operatorname{do}(X_D)\right).$$

Note that this Markov kernel is only dependent on x_C and x_D, but constant in $x_{J\setminus D}$. Such a Markov kernel is unique up to $P_{\mathcal{M}}\left(X_B, X_C | \operatorname{do}(X_{D\cup J})\right)$-null sets.

2. Action/observation exchange*: If we have:*

$$A \overset{\sigma}{\underset{G_{\operatorname{do}(I_B, D)}}{\perp}} I_B \,|\, B \cup C \cup D,$$

then there exists a Markov kernel:

$$P\left(X_A | \cancel{\operatorname{do}}(X_B), X_C, \operatorname{do}(X_{D\cup J})\right) : \mathcal{X}_B \times \mathcal{X}_C \times \mathcal{X}_D \dashrightarrow \mathcal{X}_A$$

that is a version of:

$$P_{\mathcal{M}}\left(X_A | X_{B_1}, \operatorname{do}(X_{B_2}), X_C, \operatorname{do}(X_{D\cup J})\right),$$

for every decomposition $B = B_1 \dot\cup B_2$ simultaneously. In short we could write:

$$P_{\mathcal{M}}\left(X_A | \operatorname{do}(X_B), X_C, \operatorname{do}(X_D)\right) = P_{\mathcal{M}}\left(X_A | X_B, X_C, \operatorname{do}(X_D)\right).$$

since these Markov kernels are constant in $x_{J\setminus D}$. Such a Markov kernel is unique up to $P_{\mathcal{M}}\left(X_B, X_C | \operatorname{do}(X_{I_B}, X_{D\cup J})\right)$-null set.

3. Insertion/deletion of action*: If we have:*

$$A \overset{\sigma}{\underset{G_{\operatorname{do}(I_B, D)}}{\perp}} I_B \,|\, C \cup D,$$

then there exists a Markov kernel:

$$P\left(X_A | \underline{\operatorname{do}(\cancel{X_B})}, X_C, \operatorname{do}(X_{D\cup J})\right) : \mathcal{X}_C \times \mathcal{X}_D \dashrightarrow \mathcal{X}_A$$

that is a version of:

$$P_{\mathcal{M}}\left(X_A | \operatorname{do}(X_{B_2}), X_C, \operatorname{do}(X_{D\cup J})\right)$$

for every subset $B_2 \subseteq B$ simultaneously. In short we could write:

$$P_{\mathcal{M}}\left(X_A | \operatorname{do}(X_B), X_C, \operatorname{do}(X_{D\cup J})\right) = P_{\mathcal{M}}\left(X_A | X_C, \operatorname{do}(X_D)\right).$$

Note that this Markov kernel is only dependent on x_C and x_D, but constant in $x_{J\setminus D}$. Such a Markov kernel is unique up to $P_{\mathcal{M}}\left(X_C | \operatorname{do}(X_{I_B}, X_{D\cup J})\right)$-null set.

Proof. The proof is also analogous to that of Corollary 4.4.2, except that it applies the global Markov property for simple SCMs, Corollary 6.8.1, instead of the one for causal Bayesian networks, Theorem 4.2.1.

□

Thus, by making use of the do-calculus, we can in some cases derive the existence of Markov kernels $P_{\mathcal{M}}(X_A \,|\, \mathrm{do}(X_B))$ where $J \not\subseteq B$; these should be interpreted as Markov kernels $\mathcal{X}_{J \cup B} \to \mathcal{X}_A$ that are constant in $x_{J \setminus B}$. Essentially, the do-calculus follows immediately from application of the global Markov property. We do not know if this do-calculus is complete in our setting: in some cases, the global Markov property could perhaps be used directly to obtain stronger conclusions.

Remark 6.9.2. *In case you are wondering how to use rule 3 to insert or delete an action for an exogenous input variable $j \in J$: this can be done via the special case $B = I_B = \emptyset$ and $D = J \setminus \{j\}$.*

6.10. Adjustment

Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ be a simple SCM. Let us assume $J \subseteq D$. We are interested in estimating the conditional causal effect:

$$P_{\mathcal{M}}(X_A | X_C, \mathrm{do}(X_B, X_D)),$$

but we only have data from:

$$P_{\mathcal{M}}(X_A, X_B, X_F | X_C, \mathrm{do}(X_D)).$$

The following (pairwise disjoint) index sets will have the following roles:

A : the outcome variables of interest.

B : the treatment or intervention variables.

C : general conditional (context) variables under which the data was collected.

D : general interventional (context) variables that were set by the experimenter, $J \subseteq D$.

F_0 : core adjustment variables, i.e. features that were measured.

F_1 : additional measured adjustment variables, with $F = F_0 \cup F_1$.

H : additional unobserved variables.

We will make use of the same extended SCM $\tilde{\mathcal{M}}$ with intervention variable I_B and graph $G_{\mathrm{do}(I_B)}$ as for stating the do-calculus.

Theorem 6.10.1 (General adjustment formula for simple SCMs)**.** *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ be a simple SCM with graph $G = G(\mathcal{M})$. Assume that all the following conditions hold in the graph $G_{\mathrm{do}(I_B, D)}$:*

$$(F_0 \cup H) \underset{G_{\mathrm{do}(I_B,D)}}{\overset{\sigma}{\perp}} I_B \,|(C \cup D), \tag{45}$$

$$A \underset{G_{\mathrm{do}(I_B,D)}}{\overset{\sigma}{\perp}} (F_1 \cup I_B) \,|(B \cup F_0 \cup H \cup C \cup D), \tag{46}$$

$$H \underset{G_{\mathrm{do}(I_B,D)}}{\overset{\sigma}{\perp}} B \,|(F \cup C \cup I_B \cup D). \tag{47}$$

Then we have the adjustment formula:

$$P_{\mathcal{M}}(X_A|X_C, \mathrm{do}(X_B, X_D)) = P_{\mathcal{M}}(X_A|X_B, X_C, X_F, \mathrm{do}(X_D)) \circ P_{\mathcal{M}}(X_F|X_C, \mathrm{do}(X_D)) \qquad a.s.$$

Proof. Analogous to that of Theorem 4.4.3, but now using the global Markov property for simple SCMs, Corollary 6.8.1, instead of the one for causal Bayesian networks, Theorem 4.2.1. □

Note that the "a.s." qualifier is imprecise, a more precise statement could be obtained by tracing the uniqueness of the various Markov kernels appearing in the proof.

The following is just the special case $F_1 = H = \emptyset$.

Corollary 6.10.2 (Conditional backdoor covariate adjustment formula for simple SCMs)**.** *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM with graph $G = G(\mathcal{M})$. Assume that the* conditional backdoor criterion *in the graph $G_{\mathrm{do}(I_B,D)}$ holds:*

1. $F \overset{\sigma}{\underset{G_{\mathrm{do}(I_B,D)}}{\perp}} I_B \,|(C \cup D)$, *and:*
2. $A \overset{\sigma}{\underset{G_{\mathrm{do}(I_B,D)}}{\perp}} I_B \,|(B \cup F \cup C \cup D)$.

Then we have the adjustment formula:

$$P_{\mathcal{M}}(X_A|X_C, \mathrm{do}(X_B, X_D)) = P_{\mathcal{M}}(X_A|X_B, X_C, X_F, \mathrm{do}(X_D)) \circ P_{\mathcal{M}}(X_F|X_C, \mathrm{do}(X_D)) \qquad a.s.$$

The following is just the even more special case $C = D = J = \emptyset$.

Corollary 6.10.3 (Backdoor covariate adjustment for simple SCMs)**.** *Let the situation be like in theorem 4.4.4 with $C = D = J = \emptyset$. Assume that the* backdoor criterion *holds:*

1. $F \overset{\sigma}{\underset{G_{\mathrm{do}(I_B)}}{\perp}} I_B$, *and:*
2. $A \overset{\sigma}{\underset{G_{\mathrm{do}(I_B)}}{\perp}} I_B \,|(B \cup F)$.

Then we have the adjustment formula:

$$P_{\mathcal{M}}(X_A|\,\mathrm{do}(X_B)) = P_{\mathcal{M}}(X_A|X_B, X_F) \circ P_{\mathcal{M}}(X_F) \qquad a.s.$$

6.11. Some 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 often be causally modelled by an SCM [BM18].

For illustration purposes we provide two examples, the first consisting of interacting masses that are attached to springs that can at equilibrium be described with a simple SCM, 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.

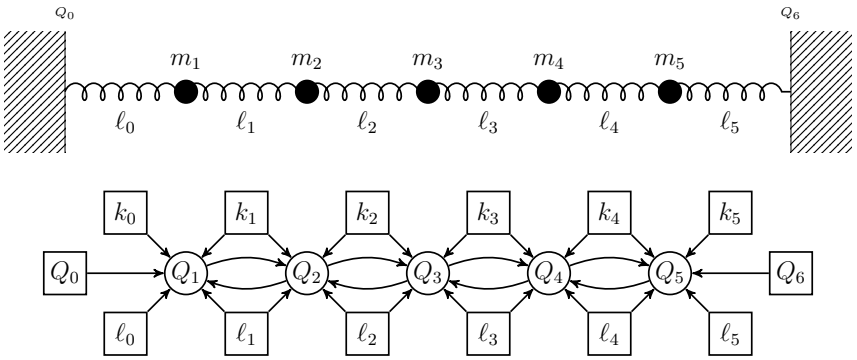


Figure 6: Damped coupled harmonic oscillator (top) and the graph of the SCM that describes the positions of the masses at equilibrium (bottom) of Example 6.11.1 for $d = 5$, where the spring lengths and constants are considered as exogenous input variables.

Example 6.11.1 (Damped coupled harmonic oscillator)**.** *Consider a one-dimensional system of d masses $m_i \in \mathbb{R}$ $(i = 1, \dots, d)$ with positions Q_i. The masses are coupled by springs, with spring constants $k_i > 0$ $(i = 0, \dots, d)$ and equilibrium lengths $\ell_i > 0$ $(i = 0, \dots, d-1)$, under influence of friction with friction coefficients $b_i > 0$ $(i = 1, \dots, d)$. The endpoints are considered fixed at positions $Q_0 < Q_{d+1}$ (see Figure 6 (top)). From elementary physics, we know that the equations of motion of this system are provided by the following differential equations*

$$\frac{d^2 Q_i}{dt^2} = \frac{k_i}{m_i}(Q_{i+1} - Q_i - \ell_i) + \frac{k_{i-1}}{m_i}(Q_{i-1} - Q_i + \ell_{i-1}) - \frac{b_i}{m_i}\frac{dQ_i}{dt} \qquad i = 1, \dots, d.$$

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 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 $\frac{d^2 Q_i}{dt^2}$ of the masses vanish (i.e., $\frac{dQ_i}{dt}, \frac{d^2 Q_i}{dt^2} \to 0$), 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 or random) variables, and the Q_i $(i = 1, \dots, d)$ as endogenous variables, we arrive at an SCM with causal

mechanism

$$f_i(q) = \frac{k_i(q_{i+1} - \ell_i) + k_{i-1}(q_{i-1} + \ell_{i-1})}{k_i + k_{i-1}}.$$

for $i = 1, \ldots, d$. Its graph is depicted in Figure 6 (bottom). This SCM 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 $0 \leq \xi_j \leq L$, the equilibrium positions of the masses correspond to the solutions of the intervened model $\mathcal{M}_{\mathrm{do}(\{j\},\xi_j)}$.

Exercise 6.11.2. *Prove that the SCM that describes the equilibrium states of a damped coupled harmonic oscillator is simple (see also Proposition 6.1.3). Hint: you can use 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 the well-known market equilibrium model from economics, can be described by a (non-simple) SCM. This example illustrates how self-cycles enrich the class of SCMs.

Example 6.11.3 (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:*

$$\begin{aligned} D &= \beta_D R + E_D \\ S &= \beta_S R + E_S \\ \frac{dR}{dt} &= D - S, \end{aligned}$$

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 the situation known as a "market equilibrium", the price is determined implicitly by the condition that demanded and supplied quantities should be equal, since $\frac{dR}{dt} = 0$ at equilibrium. At equilibrium, hence, we obtain an SCM $\mathcal{M}$ with causal mechanism defined by:

$$\begin{aligned} f_D(d, s, r, e_D, e_S) &:= \beta_D r + e_D \\ f_S(d, s, r, e_D, e_S) &:= \beta_S r + e_S \\ f_R(d, s, r, e_D, e_S) &:= r + (d - s). \end{aligned}$$

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 7 (left).

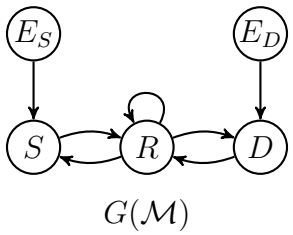


Figure 7: The graph of the SCM $\mathcal{M}$ of Example 6.11.3.

Exercise 6.11.4. *Prove that the SCM $\mathcal{M}$ that describes the equilibrium states of the price-supply-demand model is 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 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 market 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 SCMs for this lecture. Generalizations of the theory presented here for simple SCMs to non-simple ones are provided in [BFPM21].

7. Causal Discovery

So far, we always assumed that an SCM was fully specified, and derived theory to draw conclusions from the SCM. For example, the do-calculus provides precise relationships between certain Markov kernels induced by the SCM. This enables us to perform *causal reasoning*.

However, we often do not have sufficient information on the causal mechanism that we are modeling to completely specify an SCM. For example, we may only know what the observed variables are, but not what the graph of the SCM 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 but perhaps unavoidably, are typically untestable).

In the rest of this chapter we will focus on the question of how to deduce partial knowledge about the observable graph from given Markov kernels. This is often called *causal discovery*. In the next chapter, we will go one step further, and replace the deduction of graphical properties from Markov kernels by the inference of graphical properties from data, i.e., we replace Markov kernels by finite samples. This will open up an entire realm of statistical issues. For example, one might study the properties of different estimators of causal effects. Estimating the causal effect of some variables on others is often called *causal inference* (although "inference" is often interpreted much broader as drawing conclusions from data and prior beliefs).

In this lecture, we will make use of the SCM formalism, but similar results can be obtained in the IL-CBN formalism.

7.1. Detecting Causal Relations

We start by formalizing Definition 1.2.3 for simple SCMs.

Definition 7.1.1. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM with graph $G = G(\mathcal{M})$. If $a \in \mathrm{Anc}^G(b)$ for $a, b \in J \cup V \cup W$ with $a \neq b$ we say that a* is a cause of b according to $\mathcal{M}$.

With the help of the do-calculus, we can now tie this notion to practical procedures to deduce the presence of causal relations from the (interventional) Markov kernels of the SCM. The following proposition expresses that only if a causes b according to $\mathcal{M}$ can a hard intervention on a have an effect on b. This formalizes our intuitive notion of what it means for a variable to cause another variable.

Proposition 7.1.2. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM with graph $G = G(\mathcal{M})$. Let $a \in V \cup W \cup J$ and $b \in V$. If $a \notin \mathrm{Anc}^G(b)$, then:*

$$P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_{J\cup\{a\}})) = P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_{J\setminus\{a\}})).$$

Proof. Assume that $a \notin \mathrm{Anc}^G(b)$.

Let us first consider the case $a \in V \cup W$. We will show that

$$b \underset{G_{\mathrm{do}(I_a)}}{\overset{\sigma}{\perp}} I_a \mid J \setminus \{a\}.$$

Assume on the contrary that there exists a walk in $G_{\mathrm{do}(I_a)}$ between b and $I_a \cup J$ that is σ-open given $J \setminus \{a\}$. It cannot contain a node from $J \setminus \{a\}$, 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 $I_a \to a \to \ldots b$. If it were a directed walk from I_a all the way to b, then we would get a contradiction with $a \notin \mathrm{Anc}^G(b)$. Therefore, it must contain a collider. This collider must be in $J \setminus \{a\}$, which is a contradiction (since no input node can be a collider on a walk).

Now consider the case $a \in J$. In that case we do not add an intervention node. We will show that

$$b \underset{G}{\overset{\sigma}{\perp}} a \mid J \setminus \{a\}.$$

Assume on the contrary that there exists a walk in G between b and J that is σ-open given $J \setminus \{a\}$. It cannot contain a node from $J \setminus \{a\}$, 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 $a \to \ldots b$ If it were a directed walk from a all the way to b, then we would get a contradiction with $a \notin \mathrm{Anc}^G(b)$. Therefore, it must contain a collider. This collider must be in $J \setminus \{a\}$, which is a contradiction (since no input node can be a collider on a walk).

In both cases, rule 3 of the do-calculus (Corollary 6.9.1) yields that[17]

$$P_{\mathcal{M}}\left(X_b | \,\mathrm{do}(X_{J \cup \{a\}})\right) = P_{\mathcal{M}}\left(X_b | \,\mathrm{do}(X_{J \setminus \{a\}})\right).$$

JM: Idea: could we shorten the proof by looking at the marginalized graph instead? □

This leads to a practical way of detecting the presence of a causal relation.

Corollary 7.1.3. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ be a simple SCM. Let $b \in V$.*

- *For $a \in V \cup W$ with $a \neq b$, if:*
  - *there exist values $x_J \in \mathcal{X}_J$, $x_a, x'_a \in \mathcal{X}_a$ such that*

$$P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_a = x_a), \mathrm{do}(X_J = x_J)) \neq P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_a = x'_a), \mathrm{do}(X_J = x_J)),$$

  - *or there exist values $x_J \in \mathcal{X}_J$, $x_a \in \mathcal{X}_a$ such that*

$$P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_a = x_a), \mathrm{do}(X_J = x_J)) \neq P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_J = x_J)),$$

  then a is a cause of b according to $\mathcal{M}$.

[17] In the case $a \in J$, apply the rule with $B = I_B = \emptyset$.

- *For $a \in J$, if there exist values $x_{J\setminus\{a\}} \in \mathcal{X}_J$, $x_a, x'_a \in \mathcal{X}_a$ such that*

$$\begin{aligned} &P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_a = x_a), \mathrm{do}(X_{J\setminus\{a\}} = x_{J\setminus\{a\}})) \\ \neq &P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_a = x'_a), \mathrm{do}(X_{J\setminus\{a\}} = x_{J\setminus\{a\}})), \end{aligned}$$

  then a is a cause of b according to $\mathcal{M}$.

In both cases, therefore, $a \in \mathrm{Anc}^{G(\mathcal{M})}(b)$.

This condition is sufficient, but not necessary. This formalizes 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.[18]

7.2. Detecting Direct Causal Relations

Another popular notion is that of direct causation. One should keep in mind that this is always relative to some set of variables. In particular, this property is not necessarily preserved under marginalization.

Definition 7.2.1. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM with graph $G = G(\mathcal{M})$. If $a \in \mathrm{Pa}^G(b)$ for $a, b \in J \cup V \cup W$ with $a \neq b$ we say that a* is a direct cause of b w.r.t. $V \cup W \cup J$ according to $\mathcal{M}$.

Proposition 7.2.2. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM. For $a, b \in J \cup V \cup W$, we have that a is a direct cause of b w.r.t. $J \cup V \cup W$ according to $\mathcal{M}$ if and only if a is a cause of b according to $\mathcal{M}_{\mathrm{do}(V\setminus\{a,b\})}$.*

Applying Proposition 7.1.2 to the intervened SCM $\mathcal{M}_{\mathrm{do}(V\setminus\{a,b\})}$ gives similar conditions to identify the presence of direct causal relations.

Corollary 7.2.3. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM. Let $b \in V$.*

- *For $a \in V \cup W$ with $a \neq b$, if:*
  - *there exist values $x_J \in \mathcal{X}_J$, $x_{V\setminus\{a,b\}} \in \mathcal{X}_{V\setminus\{a,b\}}$, $x_a, x'_a \in \mathcal{X}_a$ such that*

$$\begin{aligned} &P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_a = x_a), \mathrm{do}(X_{V\setminus\{a,b\}}), \mathrm{do}(X_J = x_J)) \\ \neq &P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_a = x'_a), \mathrm{do}(X_{V\setminus\{a,b\}}), \mathrm{do}(X_J = x_J)), \end{aligned}$$

[18] 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.

– *or there exist values* $x_J \in \mathcal{X}_J$, $x_{V\setminus\{a,b\}} \in \mathcal{X}_{V\setminus\{a,b\}}$, $x_a \in \mathcal{X}_a$ *such that*

$$\begin{aligned} &P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_a = x_a), \mathrm{do}(X_{V\setminus\{a,b\}} = x_{V\setminus\{a,b\}}), \mathrm{do}(X_J = x_J)) \\ \neq &P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_{V\setminus\{a,b\}} = x_{V\setminus\{a,b\}}), \mathrm{do}(X_J = x_J)), \end{aligned}$$

then a *is a direct cause of* b *w.r.t.* $V \cup J \cup W$ *according to* $\mathcal{M}$.

- *For* $a \in J$, *if there exist values* $x_{J\setminus\{a\}} \in \mathcal{X}_J$, $x_{V\setminus\{a,b\}} \in \mathcal{X}_{V\setminus\{a,b\}}$, $x_a, x'_a \in \mathcal{X}_a$ *such that*

$$\begin{aligned} &P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_a = x_a), \mathrm{do}(X_{V\setminus\{a,b\}} = x_{V\setminus\{a,b\}}), \mathrm{do}(X_{J\setminus\{a\}} = x_{J\setminus\{a\}})) \\ \neq &P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_a = x'_a), \mathrm{do}(X_{V\setminus\{a,b\}} = x_{V\setminus\{a,b\}}), \mathrm{do}(X_{J\setminus\{a\}} = x_{J\setminus\{a\}})), \end{aligned}$$

then a *is a direct cause of* b *w.r.t* $V \cup J \cup W$ *according to* $\mathcal{M}$.

In both cases, therefore, $a \in \mathrm{Pa}^{G(\mathcal{M})}(b)$.

Proof. Apply the previous corollary to the intervened SCM

$$\mathcal{M}_{\mathrm{do}(V\setminus\{a,b\})} = \langle J \cup V \setminus \{a,b\}, \{a,b\}, W, \mathcal{X}, P, f_{\{a,b\}} \rangle,$$

and note that $\mathrm{Pa}^{G(\mathcal{M})}(b) \cup \{b\} = \mathrm{Anc}^{G(\mathcal{M}_{\mathrm{do}(V\setminus\{a,b\})})}(b)$. □

Note further that this method to identify a direct causal effect may not be very practical, as it requires intervening on *all* endogenous and exogenous input variables (except b) simultaneously.

7.3. Detecting Confounding

Ancestral relations signify the existence of directed paths. For graphs without bidirected edges, we can also express the existence of a trek in terms of ancestral relations.

Proposition 7.3.1. *Let* $G = \langle J, V, E \rangle$ *be a CDG. For* $a, b \in V$*: there exists a trek between nodes* a *and* b *in* G *if and only if there exists a node* $c \in V$ *such that* $c \in \mathrm{Anc}^{G_{\mathrm{do}(b)}}(a)$ *and* $c \in \mathrm{Anc}^{G_{\mathrm{do}(a)}}(b)$.

This allows us to formulate two other intuitive notions formally, that of "a confounder" and the related notion of "confounding", both for simple SCMs.

Definition 7.3.2. *Let* $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ *be a simple SCM with graph* $G = G(\mathcal{M})$. *If* $c \in \mathrm{Anc}^{G_{\mathrm{do}(b)}}(a)$ *and* $c \in \mathrm{Anc}^{G_{\mathrm{do}(a)}}(b)$ *for* $a, b \in V$ *with* $a \neq b$ *and* $c \in V \cup W$, *then we say that* c is a confounder of a and b according to $\mathcal{M}$, *and that* a and b are confounded according to $\mathcal{M}$.

Note that exogenous input nodes in J are never called confounders, by definition.[19]

To test whether some variable is a confounder of two other variables, we can simply apply the results for detecting causal relations (after appropriate interventions). To detect (possibly latent) confounding, we can make use of the following.

[19] The reason is that they do not lead to "spurious dependences" as long as one does not mix distributions for different values of X_J.

Proposition 7.3.3. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM with graph $G = G(\mathcal{M})$. Let $a \neq b \in V$. If $b \notin \mathrm{Anc}^G(a)$ and a and b are not confounded according to $\mathcal{M}$, then $P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_a), \mathrm{do}(X_J)) = P_{\mathcal{M}}(X_b \,|\, X_a, \mathrm{do}(X_J))$.*

Proof. We will show that the assumptions imply $b \perp^{\sigma}_{G_{\mathrm{do}(I_a)}} I_a \,|\{a\} \cup J$.

Suppose on the contrary that there exists a walk in $G_{\mathrm{do}(I_a)}$ between b and $I_a \cup J$ in $G_{\mathrm{do}(I_a)}$ that is $(J \cup \{a\})$-σ-open. It cannot contain nodes from J, 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 of $G_{\mathrm{do}(I_a)}$ (σ-blocking the walk). Hence it must be of the form $I_a \to a \cdots b$. Then there exists a walk $I_a \to a \cdots b$ in $G_{\mathrm{do}(I_a)}$ that is $(J \cup \{a\})$-σ-open and contains at least one collider. Indeed, suppose we have such a walk without a collider. Then it must be a directed walk $I_a \to a \to \cdots \to d \to \cdots \to b$ where $a \to \cdots \to d$ is the longest subwalk that spans a single strongly connected component of $G_{\mathrm{do}(I_a)}$ (equivalently: of G). If $d = a$, the walk would not be $(J \cup \{a\})$-σ-open. If $d = b$, we would get a contradiction with the assumption $b \notin \mathrm{Anc}^G(a)$. Therefore, d must point to a node in another strongly connected component. We can now replace the subwalk $a \to \cdots \to d$ by a directed walk through the same strongly connected component in the other direction, $a \leftarrow \cdots \leftarrow d$. In this way we obtain the walk $I_a \to a \leftarrow \cdots \leftarrow d \to \ldots b$ which is also $(J \cup \{a\})$-σ-open, and contains a as a collider.

Therefore, there must exist a walk $I_a \to a \cdots b$ in $G_{\mathrm{do}(I_a)}$ that is $(J \cup \{a\})$-σ-open and contains a collider. Each collider on the walk must be a. There must exist such a walk of minimal length, which can contain only a single collider. That walk must be of the form $I_a \to a \leftarrow \cdots b$, where the part between a and b contains no colliders and is not a directed walk from b to a. Hence it must be of the form $I_a \to a \leftarrow \cdots \leftarrow c \to \cdots \to b$, where b does not appear in the subwalk between a and c, and a does not appear in the subwalk between c and b. Contradiction. Hence $b \perp^{\sigma}_{G_{\mathrm{do}(I_a)}} I_a \,|\{a\} \cup J$.

Invoking rule 2 of the do-calculus (Corollary 6.9.1) then gives

$$P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_a), \mathrm{do}(X_J)) = P_{\mathcal{M}}(X_b \,|\, X_a, \mathrm{do}(X_J)).$$

JM: Idea: could we shorten the proof by looking at the marginalized graph instead? □

This leads to the following criterion to detect confounding:

Corollary 7.3.4. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM. Let $a, b \in V$ with $a \neq b$. If b is not a cause of a according to $\mathcal{M}$, and*

$$P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_a), \mathrm{do}(X_J)) \neq P_{\mathcal{M}}(X_b \,|\, X_a, \mathrm{do}(X_J)),$$

then a and b are confounded according to $\mathcal{M}$.[20]

This condition is sufficient, but not necessary. To apply it, we need to know already that b does not cause a according to $\mathcal{M}$. By swapping the roles of a and b, we can also

[20] The inequality of the two Markov kernels means that the one on the left is not a version of the one on the right.

use this if we know that a does not cause b. How to detect confounding if a and b are part of a causal cycle is an open research problem.

Similarly to how we applied the result for detecting a causal relation to the intervened SCM $\mathcal{M}_{\mathrm{do}(V\setminus\{a,b\})}$ in order to arrive at a criterion to detect a direct causal relation, we can apply Corollary 7.3.4 to the intervened SCM $\mathcal{M}_{\mathrm{do}(V\setminus\{a,b\})}$ to arrive at a condition to identify the presence of common direct cause.

Corollary 7.3.5. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM. Let $a \neq b \in V$.*
If b is not a direct cause of a w.r.t $J \cup V \cup W$ according to $\mathcal{M}$, and

$$P_{\mathcal{M}}(X_b \,|\, \mathrm{do}(X_a), \mathrm{do}(X_{V\setminus\{a,b\}}), \mathrm{do}(X_J)) \neq P_{\mathcal{M}}(X_b \,|\, X_a, \mathrm{do}(X_{V\setminus\{a,b\}}), \mathrm{do}(X_J))$$

then a and b have a common direct cause $c \in W$ w.r.t. $J \cup V \cup W$ according to $\mathcal{M}$.

Proof. If b is not a direct cause of a w.r.t. $J \cup V \cup W$ according to $\mathcal{M}$, then b is not cause of a according to $\mathcal{M}_{\mathrm{do}(X_{V\setminus\{a,b\}})}$, and with the assumed inequality of the Markov kernels, Corollary 7.3.4 implies that a and b are confounded according to $\mathcal{M}_{\mathrm{do}(X_{V\setminus\{a,b\}})}$. Hence there is a trek between a and b in $G_{\mathrm{do}(X_{V\setminus\{a,b\}})}$, which means that there is a node $c \in W$ such that $a \leftarrow c \rightarrow b \in G$. □

In the potential outcome literature, one encounters other criteria for unconfoundedness that are formulated in terms of counterfactuals.

Proposition 7.3.6. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f\rangle$ be a simple SCM. Let $a \neq b \in V$ and assume that $\mathcal{X}_a = \{0, 1\}$. Consider a "triplet" extension $\mathcal{M}^3$ that connects three parallel worlds, similarly to how the twin network $\mathcal{M}^{\mathrm{twin}}$ connects two parallel worlds. If b does not cause a according to $\mathcal{M}$, and a and b are not confounded according to $\mathcal{M}$, then*

$$X_a \underset{P_{\mathcal{M}^3_{\mathrm{do}(X_{a'}=0, X_{a''}=1)}}}{\perp\!\!\!\perp} \{X_{b'}, X_{b''}\}. \tag{48}$$

Proof. Suppose there exists a σ-open walk in $G(\mathcal{M}^3_{\mathrm{do}(X_{a'}=0, X_{a''}=1)}) = G(\mathcal{M}^3)_{\mathrm{do}(a',a'')}$ between a and b' or between a and b''. Then there must be such a walk of minimal length. It cannot contain any collider. It cannot be a directed walk because then it would have to pass through an exogenous random node in W, and those nodes have no incoming edges. Therefore it must be a trek of the form $a \leftarrow \cdots \leftarrow c \rightarrow \cdots \rightarrow b'$. But then $c \in W$, because only nodes in W can be ancestors of endogenous nodes in different "worlds". This implies the existence of a similar trek of the form $a \leftarrow \cdots \leftarrow c \rightarrow \cdots \rightarrow b$ in $G(\mathcal{M})$. But then a and b are confounded according to $\mathcal{M}$, contradicting the assumptions. Hence

$$a \underset{G(\mathcal{M}^3)_{\mathrm{do}(a',a'')}}{\perp} \{b', b''\}.$$

By the global Markov property we conclude that

$$X_a \underset{P_{\mathcal{M}^3_{\mathrm{do}(X_{a'}=0, X_{a''}=1)}}}{\perp\!\!\!\perp} \{X_{b'}, X_{b''}\}.$$

□

Equation (48), typically written in terms of potential outcomes as

$$X_a \perp\!\!\!\perp \{X_{b'}^{\mathrm{do}(X_{a'}=0)}, X_{b''}^{\mathrm{do}(X_{a''}=1)}\},$$

is sometimes taken as the definition of "no confounding" in the potential outcome framework when the task is to estimate the causal effect of a on b. It is a weaker assumption since it only requires a joint distribution on the various potential outcomes to be defined, and does not presuppose the existence of an underlying SCM that induces these distributions via the triple-twin operation.

7.4. Ignoring Details through Marginalizations

The graphical marginalization operation preserves ancestral relations and the existence of treks (which can be thought of as a graphical notion of confounding).

Proposition 7.4.1. *Let $G = \langle J, V, E, L\rangle$ be a CDMG, and let $U \subseteq V$. For $a, b \in J \cup V$ with $\{a, b\} \cap U = \emptyset$:*

1. *there is a directed path from a to b in G if and only if there is a directed path from a to b in $G^{\setminus U}$;*
2. *there is a trek between a and b in G if and only if there is a trek between a and b in $G^{\setminus U}$.*

Proof. 1. That is part of Remark 3.2.2.

2. Let

$$v = v_0 \tlarrow v_1 \tlarrow \cdots \tlarrow v_{k-1} \tlarrow\!\!* v_k \tarrow \cdots \tarrow v_{n-1} \tarrow v_n = w,$$

be a trek in G. If one marginalizes out a single node ℓ that is not on the trek, then the same trek exists in $G^{\setminus\{\ell\}}$. If one marginalizes out a single node ℓ on the trek (except for the endpoints) one obtains again a trek in $G^{\setminus\{\ell\}}$. The statement follows by induction.

□

Remark 7.4.2. *As special cases, we obtain:*

1. *there is a directed path from a to b in G if and only if $a \to b$ is present in $G^{\setminus(V\setminus\{a,b\})}$;*
2. *there is a trek between a and b in G if and only if $a \leftrightarrow b$ in $G^{\setminus(V\setminus\{a,b\})}$.*

We often deal with the situation that only part of the variables are observed, and others are latent.

Notation 7.4.3. *For a simple SCM $\mathcal{M}$, let $O \subseteq V \cup W$ be the set of observed variables, where we will always assume J to be observed as well. The corresponding marginalized graph $(G(\mathcal{M}))^{\setminus((V\cup W)\setminus O)}$ is called the* observable *graph of $\mathcal{M}$ and also denoted as $G^{J\cup O}(\mathcal{M})$. It has nodes $J \cup O$. Similarly, the corresponding observational Markov kernel is $P_{\mathcal{M}}(X_O \mid \mathrm{do}(X_J))$, and the corresponding interventional Markov kernels are $P_{\mathcal{M}}(X_{O\setminus T} \mid \mathrm{do}(X_{J\cup T}))$ for $T \subseteq O$.*

We formulated the global Markov property for simple SCMs in terms of its graph $G(\mathcal{M})$. What to do if we only know the observable graph $G^{J\cup O}(\mathcal{M})$?

Remark 7.4.4. *Lemma 3.3.6 shows that for d/σ-separation statements concerning only observed variables, the observable graph $G^{J\cup O}(\mathcal{M})$ contains all information we need. In other words, the observable graph hides irrelevant details if one is only interested in separations of the observed variables. Proposition 7.4.1 shows that the observable graph $G^{J\cup O}(\mathcal{M})$ also contains all information on whether observed variables are causally related according to $\mathcal{M}$, or whether two observed endogenous variables are confounded according to $\mathcal{M}$. Summarizing: the graphical marginalization preserves ancestral relations, treks and separations.*

The probabilistic marginalization preserves the notion of conditional independence. I.e., for $A, B, C \subseteq J \cup O$, we have

$$X_A \underset{P_{\mathcal{M}}(X_V, X_W | X_J)}{\perp\!\!\!\perp} X_B \mid X_C \iff X_A \underset{P_{\mathcal{M}}(X_O | X_J)}{\perp\!\!\!\perp} X_B \mid X_C.$$

These notions also interact nicely, e.g., $\mathcal{M}$ is σ/d-faithful w.r.t. O if and only if $\mathcal{M}_O$ is σ/d-faithful.

These properties taken together are very powerful, as they allow us to ignore latent details when performing causal reasoning.

The following exercise illustrates that we do not need to know the latent structure in order to derive conclusions regarding observed variables and events.

Exercise 7.4.5. *Reichenbach's Principle of Common Cause states that if two events are dependent, then one must cause the other or the events are confounded (or any combination of these three possibilities). We can make this precise for simple SCMs in the following way. Assume that $\mathcal{M}$ is a simple SCM with two observed endogenous variables X, Y (and possibly other latent variables as well, but no exogenous input nodes). Prove that $X \not\perp\!\!\!\perp_{P_{\mathcal{M}}} Y$ implies that $X \to Y$, $X \leftarrow Y$ or $X \leftrightarrow Y$ in $G^{\{X,Y\}}(\mathcal{M})$.*

7.5. Randomized Controlled Trials

Randomized controlled trials, also known as A/B-testing in engineering, provide the gold standard to discover causal relations and to estimate the 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.[21] Patients are treated depending on the assigned value of C, i.e., patients in the treatment group are treated

[21] 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.

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 SCMs. 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 SCM 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 three possibilities: exogenous input, exogenous random, and endogenous. We will discuss each of these three possibilities in sequence.

Let us start by considering the treatment variable C as an exogenous input variable. We are interested in answering two questions. The first is "Does treatment cause outcome?", where we interpret this question as that the hypothetical causal relation should hold according to the underlying SCM $\mathcal{M}$. In terms of the observable graph $G^{\{C,X\}}(\mathcal{M})$, this is then equivalent to asking "Is $C \to X$ in $G^{\{C,X\}}(\mathcal{M})$?". The second question is "What is the causal effect of treatment on outcome?". We interpret this as asking for the Markov kernel $P_{\mathcal{M}}(X \mid \mathrm{do}(C))$.

Proposition 7.5.1. *Let $\mathcal{M}$ be a simple SCM with a single exogenous input variable C and an endogenous variable X, both of which are observed (and possibly other latent variables as well). A dependence*

$$X \underset{P_{\mathcal{M}}(X|\mathrm{do}(C))}{\not\perp\!\!\!\perp} C \tag{49}$$

implies that C causes X according to $\mathcal{M}$.

Proof. This follows immediately from Proposition 7.1.2. □

Alternatively, we can consider the treatment variable as an exogenous *random* variable.

Proposition 7.5.2. *Let $\bar{\mathcal{M}}$ be a simple SCM with an exogenous random variable C and an endogenous variable X, both of which are observed (and possibly other latent variables as well). A dependence*

$$X \underset{P_{\bar{\mathcal{M}}}(X,C)}{\not\perp\!\!\!\perp} C \tag{50}$$

implies that C causes X according to $\bar{\mathcal{M}}$. The causal effect of C on X satisfies:

$$P_{\bar{\mathcal{M}}}\big(X \,|\, \mathrm{do}(C)\big) \overset{a.s.}{=} P_{\bar{\mathcal{M}}}(X \,|\, C). \tag{51}$$

Proof. Denote the observable graph as $\bar{G} := G^{\{C,X\}}(\bar{\mathcal{M}})$. It has two nodes, and it either has no edge at all, in which case C does not cause X according to $\bar{\mathcal{M}}$, or it has a single edge $C \to X$, in which case C causes X according to $\bar{\mathcal{M}}$. By the Markov property (Corollary 6.8.1), if the edge $C \to X$ were absent in $\bar{G}$, then $X \perp\!\!\!\perp_{P_{\bar{\mathcal{M}}}(X,C)} C$. In both cases, the causal do-calculus applied to $\bar{G}$ yields the identity (51). □

A third option is to consider the treatment variable as *endogenous*. One situation in which this makes sense 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 the assigned treatment, (even though coin flip outcome may still be an important cause of the assigned treatment). In this modeling variant, we assume the existence of a simple SCM $\mathcal{M}$ with endogenous variables C, X, both of which are observed, and possibly other latent variables that provides an accurate model. C here retains the meaning of assigned treatment (and 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 7.5.3. *Let $\tilde{\mathcal{M}}$ be a simple SCM with two observed endogenous variables C, X (and possibly other latent variables as well). Under the following two assumptions:*

1. *X does not cause C according to $\tilde{\mathcal{M}}$, and*
2. *C and X are not confounded according to $\tilde{\mathcal{M}}$,*

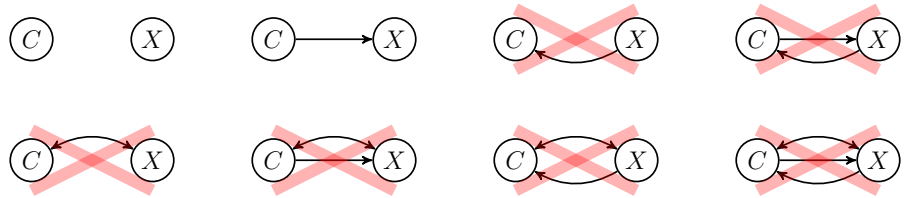
a dependence

$$X \underset{P_{\tilde{\mathcal{M}}}(X,C)}{\not\perp\!\!\!\perp} C \tag{52}$$

implies that C causes X according to $\tilde{\mathcal{M}}$, and the causal effect of C on X satisfies:

$$P_{\tilde{\mathcal{M}}}\big(X \,|\, \mathrm{do}(C)\big) \overset{a.s.}{=} P_{\tilde{\mathcal{M}}}(X \,|\, C). \tag{53}$$

Proof. Denote the observable graph as $\tilde{G} := G^{\{C,X\}}(\tilde{\mathcal{M}})$. The first assumption is equivalent to $C \leftarrow X \notin \tilde{G}$, and the second assumption is equivalent to $C \leftrightarrow X \notin \tilde{G}$. Hence, out of the eight possible graphs $\tilde{G}$, only two satisfy the assumptions:



By the Markov property, if the edge $C \to X$ were absent in $\tilde{G}$, then $X \perp\!\!\!\perp_{P_{\tilde{\mathcal{M}}}(X,C)} C$. In both cases, the causal do-calculus applied to $\tilde{G}$ yields the identity (53). □

Equation (49) is equivalent to the existence of values $c, c' \in \mathcal{X}_C$ such that

$$P_{\mathcal{M}}(X \mid \mathrm{do}(C = c)) \neq P_{\mathcal{M}}(X \mid \mathrm{do}(C = c')).$$

Equation (50) is equivalent to the existence of values

$$P_{\tilde{\mathcal{M}}}(X \mid C = c) \neq P_{\tilde{\mathcal{M}}}(X \mid C = c')$$

for every version of $P_{\tilde{\mathcal{M}}}(X \mid C)$ (and something similar holds for equation (52)). 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 SCM that provides an accurate model, in all three cases, we made the following (implicit or explicit) causal assumptions regarding the treatment variable:

1. outcome X does not cause treatment C;
2. outcome X and treatment C are not confounded, i.e., the values for the treatment variable are assigned independently of other (latent) factors that may influence the outcome.

The first assumption is commonly deemed justified if the outcome is an event that occurs later in time than the treatment event. The second assumption is usually defended by appealing to randomization. Indeed, if treatment is decided solely by a proper coin flip, then it seems reasonable to assume that there is no irreducible common cause of coin flip and outcome (where "irreducible" means that it cannot be separated into statistically independent separate causes of both).

Exercise 7.5.4. *Think like a conspiracy theorist and imagine situations in which the first assumption is* not *valid. Do the same for the second assumption.*

Another implicit assumption we made is that the data was not subject to selection bias. In other words, no data is missing (except perhaps completely at random). For example, if patients that suffer from certain treatment side effects were removed from the data set, then this assumption may be violated.

We have shown (in three 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 outcome, at least in this population of patients. Supposing that treatment is completely randomized, there are two conceptually different ways of testing this in the data, depending on whether we treat the data as a single pooled data set, or rather as two separate data sets (each one corresponding to a particular patient group), see also Figure 8. To test whether $P(X \mid \mathrm{do}(C = 0)) \neq P(X \mid \mathrm{do}(C = 1))$, we can test whether the distribution of X is statistically different in the two groups. This can be tested with a two-sample test, for example, a t-test or a Wilcoxon test. The other alternative is to

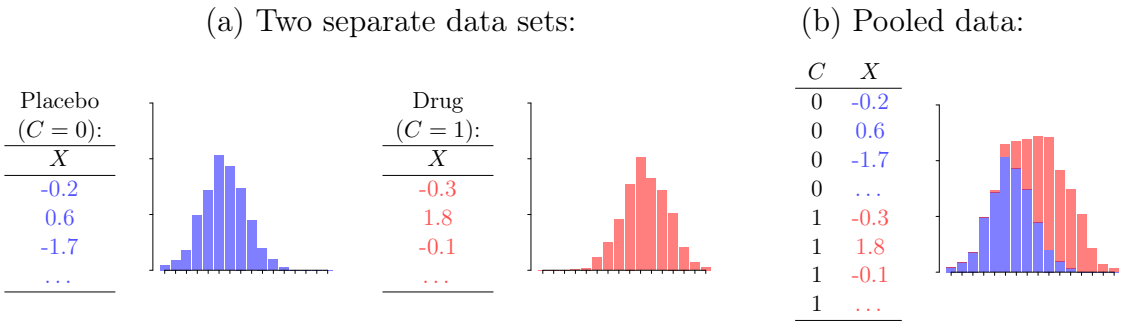


Figure 8: Illustration of the data from an example randomized controlled trial. When treatment C is randomized, the data can either be interpreted as (a) two separate data sets, one for the treatment and one for the control group, or (b) as a single data set including a context variable indicating treatment/control. Note that in this particular example, C is dependent on X in the pooled data (or equivalently, the distribution of X differs between contexts $C = 0$ and $C = 1$), which implies that C is a cause of X.

consider the data as a single *pooled* data set. The question then becomes whether the conditional distribution of X given $C = 0$ differs from the conditional distribution of X given $C = 1$, i.e., whether $P(X \mid C = 0) \neq P(X \mid C = 1)$. This can be done with a conditional independence test, for example, Fisher's exact test, or a partial correlation test. In the next lecture, we will look in more detail at possible tests.

In the end, the three ways of modeling the RCT are only slightly different. If one formally considers treatment as an exogenous input variable, but then also assumes that its values are randomly assigned, then the differences are purely cosmetical. However, there is one advantage that the exogenous input approach has over the other two: here we do not model *at all* how the values of treatment are chosen (except for the exogeneity assumptions). This allows more freedom in the experimental design and sampling scheme design. For example, one can decide ahead of 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 end up with exactly the same number of patients in both groups.

A fourth way to formalize the randomized controlled setting is by using potential outcomes. For a binary treatment variable, we introduce 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 $X_n := X_n^{\mathrm{do}(C_n)}$. In practice, we only observe the actual outcome, and the other potential outcome remains latent. The task of estimating the causal effect of treatment on outcome is then often formulated as estimating the *average treatment effect (ATE)*

$$\tau := \mathbb{E}(X_n^{\mathrm{do}(C_n=1)} - X_n^{\mathrm{do}(C_n=0)}).$$

To do so, one assumes that C_n is randomized. This motivates the assumption that

treatment and outcome are unconfounded, i.e., with Proposition 7.3.6:

$$\{X_n^{\mathrm{do}(C_n=1)}, X_n^{\mathrm{do}(C_n=0)}\} \perp\!\!\!\perp C_n.$$

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 a fundamentally unobservable quantity; however, the average treatment effect can be estimated from observed data. In the SCM setting, we can think of the potential outcomes as counterfactuals in a twin SCM. However, when assuming an underlying SCM, there is no need to go to the counterfactual level, as one can simply define the ATE as

$$\tau := \mathbb{E}_{\mathcal{M}}(X \,|\, \mathrm{do}(C = 1)) - \mathbb{E}_{\mathcal{M}}(X \,|\, \mathrm{do}(C = 0)).$$

In the presence of observed covariates Z, one often considers also the *conditional average treatment effect (CATE)*, which we can define as

$$\mathbb{E}_{\mathcal{M}}(X \,|\, \mathrm{do}(C = 1), Z) - \mathbb{E}_{\mathcal{M}}(X \,|\, \mathrm{do}(C = 0), Z).$$

when assuming an underlying SCM. 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].

7.6. Faithfulness

The converse statement of the global Markov property for simple SCMs (Corollary 6.8.1) and for causal Bayesian networks (Theorem 4.2.1) is called "faithfulness".

Definition 7.6.1. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ be a simple SCM with graph $G(\mathcal{M})$ and observational Markov kernel $P_{\mathcal{M}}(X_V, X_W \,|\, \mathrm{do}(X_J))$. $\mathcal{M}$ is called* σ-faithful *if for all $A, B, C \subseteq J \cup V \cup W$ (not necessarily disjoint):*

$$A \overset{\sigma}{\underset{G(\mathcal{M})}{\perp}} B \,|\, C \implies X_A \underset{P_{\mathcal{M}}(X_V, X_W \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C \tag{54}$$

It is called d-faithful *if for all $A, B, C \subseteq J \cup V \cup W$ (not necessarily disjoint):*

$$A \overset{d}{\underset{G(\mathcal{M})}{\perp}} B \,|\, C \implies X_A \underset{P_{\mathcal{M}}(X_V, X_W \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C \tag{55}$$

For a subset $O \subseteq V \cup W$, we say that $\mathcal{M}$ is σ-faithful w.r.t. O *if (54) 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 SCM is called σ-faithful (d-faithful) if each conditional independence in the induced Markov kernel is due to a σ-separation (d-separation).

Faithfulness may fail for various reasons:

- Deterministic relationships may lead to additional conditional independences, but are not exploited to the Markov property;
- Effects may cancel out;
- If cycles are present, and (i) all variables are discrete, or (ii) interactions are linear;
- 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 7.6.2. *Take an SCM 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 $Y \not\perp\!\!\!\perp Z$ but $Y \perp\!\!\!\perp Z$. This simple (even acyclic) SCM is not faithful due to X being a constant.

The next example illustrates how canceling effect may lead to a faithfulness violation.

Example 7.6.3. *Take an SCM 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 $X \not\perp\!\!\!\perp Z$ but $X \perp\!\!\!\perp Z$.

One can show that in certain special cases, the global Markov property in terms of d-separation even holds for simple SCMs.

Proposition 7.6.4. *Let $\mathcal{M} = \langle J, V, W, \mathcal{X}, P, f \rangle$ be a simple SCM with graph $G(\mathcal{M})$ and observational Markov kernel $P_{\mathcal{M}}(X_V, X_W \,|\, \mathrm{do}(X_J))$. If $J = \emptyset$ and one of the three conditions applies:*

1. *all spaces $\mathcal{X}_v$ with $v \in V$ are discrete, or*
2. *the causal mechanism f is affine and the exogenous distribution has a density w.r.t. Lebesgue measure, or*
3. *$\mathcal{M}$ is acyclic,*

then for all $A, B, C \subseteq J \cup V \cup W$ (not necessarily disjoint):

$$A \overset{d}{\underset{G(\mathcal{M})}{\perp}} B \,|\, C \implies X_A \underset{P_{\mathcal{M}}(X_V, X_W \,|\, \mathrm{do}(X_J))}{\perp\!\!\!\perp} X_B \,|\, X_C$$

The proofs are given in [FM17].

7.7. 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 which 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 SCMs):

Proposition 7.7.1. *Let $\mathcal{M}$ be a simple SCM with observed endogenous variables $O = \{1, 2, 3\} \subseteq V$. Suppose that it is σ-faithful (w.r.t. O). If X_2 is not a cause of X_1 according to $\mathcal{M}$, the following conditional (in)dependencies*[22] *in the observational distribution $P_{\mathcal{M}}(X_1, X_2, X_3)$*

$$X_1 \not\perp\!\!\!\perp X_2, \quad X_2 \not\perp\!\!\!\perp X_3, \quad X_1 \perp\!\!\!\perp X_3 \,|\, X_2$$

imply that the observable graph $G^O(\mathcal{M})$ must be one of the three DMGs in Figure 9. Hence,

1. *X_3 is not a cause of X_2 according to $\mathcal{M}$;*
2. *X_2 is a direct cause of X_3 w.r.t $\{1, 2, 3\}$ according to $\mathcal{M}$;*
3. *X_2 and X_3 are not confounded according to $\mathcal{M}$;*
4. *the causal effect of X_2 on X_3 is given by:*

$$P_{\mathcal{M}}\big(X_3 \,|\, \mathrm{do}(X_2)\big) \overset{a.s.}{=} P_{\mathcal{M}}(X_3 \,|\, X_2). \tag{56}$$

Proof. The proof proceeds by enumerating all (possibly cyclic) DMGs on three variables that the observable graph $G^O(\mathcal{M})$ 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 implies that there is no directed edge $X_2 \to X_1$ in the graph $\mathcal{G}^O(\mathcal{M})$. If there were an edge between X_1 and X_3, $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$ would not hold (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 (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 have an arrowhead at 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 graphs are the ones in Figure 9. The causal do-calculus applied to $G^O(\mathcal{M})$ yields (56). □

[22] Henceforth, we will no longer always write explicitly the Markov kernel as a subscript to the conditional independence symbol if it is clear from the context which Markov kernel is meant.

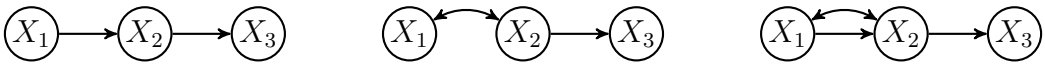


Figure 9: All possible observable graphs detected by LCD.

LCD has been applied to infer signalling networks from mass-cytometry data. A high-dimensional adaptation has been shown to be successful in predicting the effects of gene knockout on gene expression levels from large-scale interventional yeast gene expression data.

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), i.e., 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 gives rise to an algorithm known as Invariant Causal Prediction.

7.8. 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. We here also give the generalization of Y-structure pattern to simple SCMs.

Proposition 7.8.1. *Let $\mathcal{M}$ be a simple SCM with observed endogenous variables $O = \{1,2,3,4\} \subseteq V$. Suppose that it is σ-faithful (w.r.t. O). The following conditional (in)dependencies in the observational distribution $P_{\mathcal{M}}(X_1, X_2, X_3, X_4)$*

$$\begin{array}{lll} X_1 \not\perp\!\!\!\perp X_4, & X_2 \not\perp\!\!\!\perp X_4, & X_1 \perp\!\!\!\perp X_2, \\ X_1 \perp\!\!\!\perp X_4 \,|\, X_3, & X_2 \perp\!\!\!\perp X_4 \,|\, X_3, & X_1 \not\perp\!\!\!\perp X_2 \,|\, X_3, \end{array}$$

imply that the observable graph $G^O(\mathcal{M})$ must be one of the nine DMGs in Figure 10. Hence,

1. *X_3 is a direct cause of X_4 w.r.t. $\{1,2,3,4\}$ according to $\mathcal{M}$;*
2. *X_3 and X_4 are unconfounded according to $\mathcal{M}$;*
3. *the causal effect of X_3 on X_4 satisfies:*

$$P_{\mathcal{M}}\big(X_4 \,|\, \mathrm{do}(X_3)\big) \overset{a.s.}{=} P_{\mathcal{M}}(X_4 \,|\, X_3). \tag{57}$$

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 10. The statements now follow. □

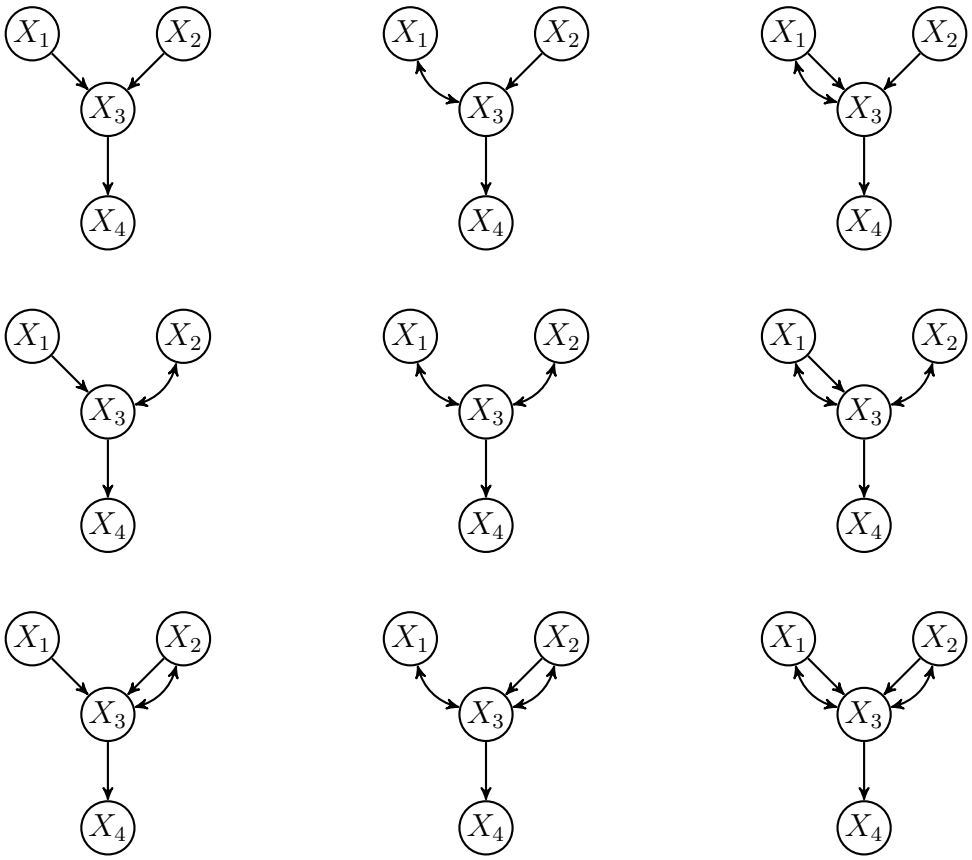


Figure 10: Observed causal graphs satisfying the "Y-structure" pattern on four variables.

This example illustrates how conditional independence patterns in the observational distribution allow one to infer certain features of the underlying causal model. This 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.

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 confounder, or both. 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 CDMGs that induce the same separations. Another 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.

8. 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?
- ... a random variable is independent of a non-random variable?
- ... a random variable is conditionally independent of a non-random variable, given another random variable and another non-random variable?

We will consider these questions only for the special case of discrete 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 random and non-random variables. We will state conditions under which the tests are asymptotically valid and consistent.

8.1. 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 $\theta \in \Theta$ 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:

$$\Theta := \{\theta \in \prod_{(x,y) \in \mathcal{X} \times \mathcal{Y}} [0, 1] : \theta_{++} = 1\}.$$

With Remark 2.5.6, 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

$$\Theta_0 := \{\theta \in \Theta : \theta_{xy} = \theta_{x+}\theta_{+y} \ \forall x \in \mathcal{X}, y \in \mathcal{Y}\} \subseteq \Theta.$$

We can also write the null hypothesis as $H_0 : \theta \in \Theta_0$. As alternative hypothesis we take that of dependence, i.e., $H_1 : X \not\perp\!\!\!\perp Y$, or equivalently, $H_1 : \theta \in \Theta_1$ with $\Theta_1 := \Theta \setminus \Theta_0$.

Suppose now that we have independent and identically distributed data $(X_n, Y_n)_{n=1}^N$ with $(X_n, Y_n) \sim P(X, Y \mid \theta)$ 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

$$P((X_n, Y_n)_{n=1}^N \mid \theta) = \bigotimes_{n=1}^N P(X = X_n, Y = Y_n \mid \theta).$$

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 $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta \setminus \Theta_0$ is as follows. One comes up with a *test statistic* $T(D)$, which is a function of the data $D \sim P(D \mid \theta)$, whose value should help us distinguish between the two hypotheses. Then one chooses a particular p-value threshold $\alpha \in (0, 1)$. 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

$$p(d) := \sup_{\theta \in \Theta_0} P(T(D) \geq T(d) \mid \theta).$$

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 Type I error (i.e., the error of incorrectly rejecting the null hypothesis), which can be achieved by choosing α sufficiently small. For causal discovery, however, we need a more symmetric treatment of the two hypotheses, as there we require both Type I error and 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 of the data:

$$P((X_n, Y_n)_{n=1}^N \mid \theta) = \prod_{n=1}^N \theta_{X_n Y_n} = \prod_{x \in \mathcal{X}} \prod_{y \in \mathcal{Y}} \theta_{xy}^{N_{xy}}$$

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 $\theta \in \Theta$, we obtain the well-known maximum likelihood estimator

$$\hat{\theta}_{xy} = \frac{N_{xy}}{N},$$

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:

$$\begin{aligned} \theta \in \Theta_0 \implies P((X_n, Y_n)_{n=1}^N \mid \theta) &= \prod_{x \in \mathcal{X}} \prod_{y \in \mathcal{Y}} (\theta_{x+}\theta_{+y})^{N_{xy}} \\ &= \left(\prod_{x \in \mathcal{X}} \prod_{y \in \mathcal{Y}} \theta_{x+}^{N_{xy}} \right) \left(\prod_{x \in \mathcal{X}} \prod_{y \in \mathcal{Y}} \theta_{+y}^{N_{xy}} \right) \\ &= \left(\prod_{x \in \mathcal{X}} \theta_{x+}^{N_{x+}} \right) \left(\prod_{y \in \mathcal{Y}} \theta_{+y}^{N_{+y}} \right). \end{aligned}$$

This is just the product of two independent multinomial distributions with 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 $\hat{\theta}^0$:

$$\begin{aligned} \frac{\sup_{\theta \in \Theta_0 \cup \Theta_1} P((X_n, Y_n)_{n=1}^N \mid \theta)}{\sup_{\theta \in \Theta_0} P((X_n, Y_n)_{n=1}^N \mid \theta)} = \frac{P((X_n, Y_n)_{n=1}^N \mid \hat{\theta})}{P((X_n, Y_n)_{n=1}^N \mid \hat{\theta}^0)} &= \prod_{x \in \mathcal{X}} \prod_{y \in \mathcal{Y}} \left(\frac{\hat{\theta}_{xy}}{\hat{\theta}^0_{x+}\hat{\theta}^0_{+y}} \right)^{N_{xy}} \\ &= \prod_{x \in \mathcal{X}} \prod_{y \in \mathcal{Y}} \left(\frac{N_{xy}N}{N_{x+}N_{+y}} \right)^{N_{xy}}. \end{aligned} \tag{58}$$

The likelihood ratio test statistic is defined as 2 times the natural logarithm of this ratio:

$$G_N := 2 \log \frac{\sup_{\theta \in \Theta_0 \cup \Theta_1} P((X_n, Y_n)_{n=1}^N \mid \theta)}{\sup_{\theta \in \Theta_0} P((X_n, Y_n)_{n=1}^N \mid \theta)} = 2 \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} N_{xy} \log \frac{N_{xy} N}{N_{x+} N_{+y}}. \tag{59}$$

For finite data, counts can be zero. In this expression, one should interpret $0 \log \frac{0}{0}$ as 0.

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,[23] i.e.,

$$\theta_{xy} > 0 \qquad \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{60}$$

Proposition 8.1.1. *Under $H_0 : X \perp\!\!\!\perp Y$, and with regularity assumption (60),*

$$G_N \rightsquigarrow \chi^2_\nu$$

with $\nu = (|\mathcal{X}| - 1)(|\mathcal{Y}| - 1)$ as sample size $N \to \infty$. In words, the test statistic G_N converges in distribution[24] as $N \to \infty$ to a chi-squared distribution with ν degrees of freedom.[25]

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 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 Θ is $|\mathcal{X}||\mathcal{Y}| - 1$, while that of Θ_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 (Proposition **??**). □

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

[23] 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.

[24] 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 $P(X_n \leq x) \to P(X_\infty \leq x)$ for all $x \in \mathbb{R}$ such that $\xi \mapsto P(X_\infty \leq \xi)$ is continuous at x.

[25] 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..

inverse (i.e., the quantile function). Indeed, if $\theta \in \Theta_0$, then $P(G_N \geq \chi^2_{\nu,1-\alpha}) \to \alpha$, 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 8.1.2. *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") of the mutual information $I(X;Y)$:

$$\begin{aligned} I(\theta) &:= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \theta_{xy} \log \frac{\theta_{xy}}{\theta_{x+}\theta_{+y}} \\ &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P(X=x, Y=y) \log \frac{P(X=x,Y=y)}{P(X=x)P(Y=y)} =: I(X;Y). \end{aligned}$$

With Jensen's inequality, one can show that $I(X;Y) \geq 0$, and that $I(X;Y) = 0 \iff X \perp\!\!\!\perp Y$. Note further that the function $\Theta \to [0,\infty) : \theta \mapsto I(\theta)$ 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 8.1.3. *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 rejects $H_0 : X \perp\!\!\!\perp Y$ if $G_N \geq \tau_N$ for some given sequence of thresholds τ_N. Under the regularity assumption (60), 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 $\theta \in \Theta$. 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.[26] 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,

$$\theta \in \Theta_1 \implies P(G_N < \tau_N) \to 0.$$

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 8.1.1, which states that $G_N \rightsquigarrow \chi^2_\nu$. 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

$$|F_{G_N}(\tau_N) - F_{\chi^2_\nu}(\tau_N)| \leq \sup_{x \in \mathbb{R}} |F_{G_N}(x) - F_{\chi^2_\nu}(x)| \to 0.$$

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

$$\theta \in \Theta_0 \implies P(G_N \geq \tau_N) \to 0,$$

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 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.[27]

[26] The convergence is "almost surely", i.e., $N_{xy}/N \overset{a.s.}{\to} \theta_{xy}$ means that $P(N_{xy}/N \to \theta_{xy}) = 1$.
[27] This is referred to as the lack of "uniformly consistent" (conditional) independence tests.

8.2. Conditional Independence for Categorial 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 2.5.6, we get (because finite spaces are standard):

$$\begin{aligned} X \underset{P(X,Y,Z)}{\perp\!\!\!\perp} Y \,|\, Z &\iff P(X,Y,Z) = P(X|Z) \otimes P(Y,Z) \\ &\iff P(X,Y|Z) = P(X|Z) \otimes P(Y|Z) \quad P(Z)\text{-a.s.} \\ &\iff \forall z \in \mathcal{Z} : \big[P(Z=z) > 0 \implies \\ &\qquad\qquad P(X,Y \mid Z=z) = P(X \mid Z=z) P(Y \mid Z=z)\big] \\ &\iff \forall z \in \mathcal{Z} : \big[P(Z=z) > 0 \implies X \underset{P(X,Y|Z=z)}{\perp\!\!\!\perp} Y\big]. \end{aligned}$$

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

$$\Theta_{XY|Z} := \{\theta \in \prod_{x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}} [0,1] : \forall_{z \in \mathcal{Z}} \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \theta_{xy|z} = 1\}.$$

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

$$\frac{P(X=x, Y=y, Z=z)}{P(Z=z)} = P(X=x, Y=y \mid Z=z) = \theta_{xy|z}.$$

We also parameterize the marginal distribution $P(Z)$ in terms of parameters $(\theta_z)_{z \in \mathcal{Z}}$ which live in space

$$\Theta_Z := \{\theta \in \prod_{z \in \mathcal{Z}} [0,1] : \sum_{z \in \mathcal{Z}} \theta_z = 1\}.$$

Any joint distribution of X, Y and Z can then be parameterized as

$$P(X=x, Y=y, Z=z \mid \theta) = \theta_z \theta_{xy|z},$$

with parameter space

$$\Theta := \Theta_Z \times \Theta_{XY|Z}.$$

We formulate the null hypothesis $H_0 : X \perp\!\!\!\perp Y \,|\, Z$ 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

$$\Theta^0_{XY|Z} := \{\theta \in \Theta : \theta_{xy|z} = \theta_{x+|z}\theta_{+y|z} \ \forall x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}\}.$$

We can then also write the null hypothesis as $H_0 : \theta \in \Theta_0$ with $\Theta_0 := \Theta_Z \times \Theta^0_{XY\,|\,Z}$. As alternative hypothesis we take that of dependence, i.e., $H_1 : X \not\perp\!\!\!\perp Y \,|\, Z$, or equivalently, $H_1 : \theta \in \Theta_1$, where $\Theta_1 := \Theta_Z \times \Theta^1_{XY\,|\,Z}$ with $\Theta^1_{XY\,|\,Z} := \Theta_{XY\,|\,Z} \setminus \Theta^0_{XY\,|\,Z}$.

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 $n = 1, \ldots, N$, with the "true" parameter $\theta \in \Theta$ unknown. In other words, we assume for the joint distribution on the observed data

$$P((X_n, Y_n, Z_n)_{n=1}^N \mid \theta) = \bigotimes_{n=1}^N P(X = X_n, Y = Y_n, Z = Z_n \mid \theta).$$

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 of the data:

$$\begin{aligned} P((X_n, Y_n, Z_n)_{n=1}^N \mid \theta) &= \prod_{n=1}^N (\theta_{X_n, Y_n|Z_n}\theta_{Z_n}) = \prod_{x\in\mathcal{X}}\prod_{y\in\mathcal{Y}}\prod_{z\in\mathcal{Z}} (\theta_{xy|z}^{N_{xyz}}\theta_z^{N_{xyz}}) \\ &= \left(\prod_{z\in\mathcal{Z}} \theta_z^{N_{++z}}\right)\left(\prod_{z\in\mathcal{Z}}\prod_{x\in\mathcal{X}}\prod_{y\in\mathcal{Y}} \theta_{xy|z}^{N_{xyz}}\right) \end{aligned}$$

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 $\theta \in \Theta$, 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:

$$\begin{aligned} \theta \in \Theta_0 \implies P((X_n, Y_n, Z_n)_{n=1}^N \mid \theta) &= \left(\prod_{z\in\mathcal{Z}} \theta_z^{N_{++z}}\right)\prod_{z\in\mathcal{Z}}\prod_{x\in\mathcal{X}}\prod_{y\in\mathcal{Y}} (\theta_{x+|z}\theta_{+y|z})^{N_{xyz}} \\ &= \left(\prod_{z\in\mathcal{Z}} \theta_z^{N_{++z}}\right)\prod_{z\in\mathcal{Z}}\left(\prod_{x\in\mathcal{X}} \theta_{x+|z}^{N_{x+z}}\right)\left(\prod_{y\in\mathcal{Y}} \theta_{+y|z}^{N_{+yz}}\right). \end{aligned}$$

The restricted maximum likelihood estimator under H_0 is

$$\left(\hat{\theta}^0_{xy|z}, \hat{\theta}^0_z\right) = \left(\hat{\theta}^0_{x+|z}\hat{\theta}^0_{+y|z}, \hat{\theta}^0_z\right) = \left(\frac{N_{x+z}N_{+yz}}{N^2_{++z}}, \frac{N_{++z}}{N}\right).$$

The likelihood ratio is obtained by dividing the likelihood for $\hat{\theta}$ by the likelihood for $\hat{\theta}^0$:

$$\begin{aligned}\frac{\sup_{\theta\in\Theta_0\cup\Theta_1} P((X_n,Y_n,Z_n)_{n=1}^N \mid \theta)}{\sup_{\theta\in\Theta_0} P((X_n,Y_n,Z_n)_{n=1}^N \mid \theta)} &= \frac{P((X_n,Y_n,Z_n)_{n=1}^N \mid \hat{\theta})}{P((X_n,Y_n,Z_n)_{n=1}^N \mid \hat{\theta}^0)} = \prod_{z\in\mathcal{Z}}\prod_{x\in\mathcal{X}}\prod_{y\in\mathcal{Y}}\left(\frac{\hat{\theta}_{xy|z}}{\hat{\theta}^0_{x+|z}\hat{\theta}^0_{+y|z}}\right)^{N_{xyz}} \\ &= \prod_{z\in\mathcal{Z}}\prod_{x\in\mathcal{X}}\prod_{y\in\mathcal{Y}}\left(\frac{N_{xyz}N_{++z}}{N_{x+z}N_{+yz}}\right)^{N_{xyz}},\end{aligned}$$

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:

$$G_N := 2\log\frac{\sup_{\theta\in\Theta_0\cup\Theta_1} P((X_n,Y_n,Z_n)_{n=1}^N \mid \theta)}{\sup_{\theta\in\Theta_0} P((X_n,Y_n,Z_n)_{n=1}^N \mid \theta)} = 2\sum_{z\in\mathcal{Z}}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} N_{xyz}\log\frac{N_{xyz}N_{++z}}{N_{x+z}N_{+yz}}. \quad (61)$$

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.,

$$\begin{cases}\theta_z > 0 & \forall z \in \mathcal{Z}, \text{ and} \\ \theta_{xy|z} > 0 & \forall x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}.\end{cases} \quad (62)$$

Proposition 8.2.1. *Under $H_0 : X \perp\!\!\!\perp Y \mid Z$, and with regularity assumption (62)*

$$G_N \rightsquigarrow \chi^2_\nu$$

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 8.1.1. The dimensionality of Θ is $|\mathcal{Z}|(|\mathcal{X}||\mathcal{Y}|-1)+(|\mathcal{Z}|-1)$, while that of Θ_0 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 $I(X;Y|Z)$:

$$\begin{aligned} I(\theta) &:= \sum_{z\in\mathcal{Z}} \theta_z \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \theta_{xy|z} \log \frac{\theta_{xy|z}}{\theta_{x+|z}\theta_{+y|z}} \\ &= \sum_{z\in\mathcal{Z}} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P(X=x, Y=y, Z=z) \log \frac{P(X=x, Y=y|Z=z)}{P(X=x|Z=z)P(Y=y|Z=z)} \\ &=: I(X;Y|Z). \end{aligned}$$

With Jensen's inequality, one can show that $I(X;Y|Z) \geq 0$, and that $I(X;Y|Z) = 0 \iff X \perp\!\!\!\perp Y \,|\, Z$. Note further that the function $\Theta \to [0,\infty) : \theta \mapsto I(\theta)$ 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 8.2.2. *Consider an infinite sequence of G tests performed on the first N samples of an infinitely large data set $(X_n, Y_n, Z_n)_{n=1}^{\infty}$, where one rejects $H_0 : X \perp\!\!\!\perp Y \,|\, Z$ if $G_N \geq \tau_N$ for some given sequence of thresholds τ_N. Under the regularity assumption (62), 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 8.1.3.

We again apply the strong law of large numbers. Let $\theta \in \Theta$. 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 (62), $\hat{\theta}_z \overset{a.s.}{\to} \theta_z$, $\hat{\theta}_{xy|z} \overset{a.s.}{\to} \theta_{xy|z}$, $\hat{\theta}_{+x|z} \overset{a.s.}{\to} \theta_{+x|z}$, 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 8.1.3 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 8.2.1, 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 8.1.3. □

9. The Fast Causal Inference Algorithm

In this final episode, we will present the Fast Causal Inference (FCI) algorithm, one of the "classic" constraint-based causal discovery algorithms originally designed for purely observational (non-experimental) data. It has a higher complexity than LCD, but it is of special interest because it can be shown to be *complete* in a certain sense. One obtains a more informative output from FCI than from LCD about the causal relations between the variables. While FCI 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). Furthermore, it has been extended to deal with prior knowledge regarding exogeneity and unconfoundedness of context variables (of the same type that we used for modeling randomized controlled trials).

In the rest of this chapter, we will not make use of exogenous input variables, since FCI and the accompanying theory has not yet been formulated to deal with those. In the proofs, we will often use the alternative formulation 3.3.1 of σ'-open/closed, exploiting Proposition 3.3.3 which expresses the equivalence of the two notions.

9.1. Inducing paths

The output of FCI will be a Partial Ancestral Graph (PAG), a certain type of graph to be defined later. An important notion in its definition is that of "inducing" walks and paths. We define this for the σ-separation setting.

Definition 9.1.1. *Let $G = \langle V, E, L\rangle$ be directed mixed graph (DMG). An* inducing walk between two nodes $i, j \in V$ *is a walk in G between i and j on which every collider is in* $\mathrm{Anc}_G(\{i, j\})$ *JM: or in S (Zhang 2006), or in* $\mathrm{Anc}_G(S)$ *(RS 2002), and each non-endpoint non-collider on the walk only has outgoing directed edges to neighboring nodes on the walk that lie in the same strongly connected component of G. If it is a path, it is called an* inducing path between $i, j \in V$.

If two nodes are adjacent, any edge connecting the two is an inducing walk (or path) between them. Figure 11 shows some simple nontrivial examples of inducing paths.

This notion has the following important properties.

Proposition 9.1.2. *Let $G = \langle V, E, L\rangle$ be a DMG and $i, j \in V$ be distinct. Then the following are equivalent:*

(i) There is an inducing walk in G between i and j;

(ii) $i \not\perp\!\!\!\perp^{\sigma}_{G} j \mid Z$ for all $Z \subseteq V \setminus \{i, j\}$;

(iii) $i \not\perp\!\!\!\perp^{\sigma}_{G} j \mid Z$ for $Z = \mathrm{Anc}_G(\{i, j\}) \setminus \{i, j\}$.

Proof. The proof is similar to that of Theorem 4.2 in [RS02].

(i) $\implies$ (ii): Assume the existence of an inducing walk between i and j in G. Let $Z \subseteq V \setminus \{i, j\}$. Consider all walks in G between i and j with the property that all

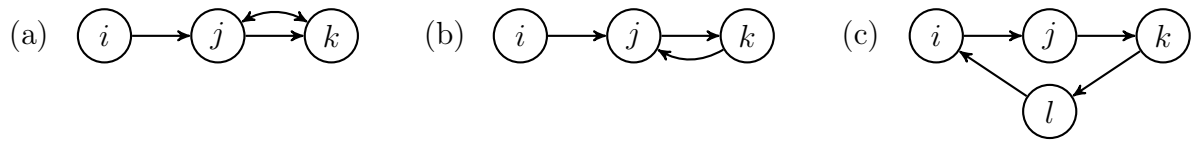


Figure 11: Three examples of non-trivial inducing paths in directed mixed graphs. (a) The walk $i \to j \leftrightarrow k$ is an inducing walk (path) between i and k. The nodes i and k cannot be σ-separated by any subset not containing i, k (indeed, $i \not\perp^{\sigma} k$ and $i \not\perp^{\sigma} k \mid j$). (b) The walk $i \to j \to k$ is an inducing walk (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 $i \not\perp^{\sigma} k \mid j$). (c) The walk $i \to j \to k$ is an inducing walk (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 $i \not\perp^{\sigma} k \mid j$).

colliders on it are in $\mathrm{Anc}_G(\{i, j\}) \cup Z$, and each non-endpoint non-collider on it is not in Z or points only to nodes in the same strongly connected component of G. Such walks exist, since the 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 $\mathrm{Anc}_G(Z)$. Suppose on the contrary the existence of a collider k on μ that is not ancestor of Z. It is either ancestor of i or of j, by assumption. Without loss of generality, assume the latter. Then there is a directed path π from k to j in G that does not pass through any node of Z. Then the subwalk of μ between i and k 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 can be extended to a walk that is σ-open given Z, by replacing each collider k on it by a walk $k \to \cdots \to z \leftarrow \cdots \leftarrow k$ for some closest descendant $z \in Z$ of k (with possibly $z = k$). Hence, i and j are σ-connected given Z.

(ii) $\implies$ (iii) is trivial.

(iii) $\implies$ (i): Suppose that i and j are σ-connected given $Z^* = \mathrm{Anc}_G(\{i, j\}) \setminus \{i, j\}$. Let π be a walk between i and j that is σ-open given Z^* and contains i and j only once. We show that π must be an inducing walk. First, all colliders on π are in $Z^* \subseteq \mathrm{Anc}_G(\{i, j\})$. Second, let k be any non-endpoint non-collider on π. Then there must be a directed subwalk 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 can only point to nodes in the same strongly connected component of G. Hence, all non-endpoint non-colliders on π can only point to nodes in the same strongly connected component of G. □

In words: there is an inducing walk between two nodes in a DMG if and only if the two nodes cannot be σ-separated by any subset of nodes that does not contain either of the two nodes. For completeness, to connect with the literature, we also formulate:

Proposition 9.1.3. *Let $G = \langle V, E, L \rangle$ be a DMG and $i, j \in V$ be distinct. Then the following are equivalent:*

(i) There is an inducing path in G between i and j;

(ii) There is an inducing walk in G between i and j.

Proof. (i) $\implies$ (ii) is trivial.

We could go via (iii) and (iv):

(iii) $i \not\perp\!\!\!\perp^{\sigma}_{G} j \,|\, Z$ for all $Z \subseteq V \setminus \{i, j\}$;

(iv) $i \not\perp\!\!\!\perp^{\sigma}_{G} j \,|\, Z$ for $Z = \mathrm{Anc}_G(\{i, j\}) \setminus \{i, j\}$,

or a direct proof by cutting out a part between repeated nodes.

(ii) $\implies$ (iii): Assume the existence of an inducing walk between i and j in G. Let $Z \subseteq V \setminus \{i, j\}$. Consider all walks in G between i and j with the property that all colliders on it are in $\mathrm{Anc}_G(\{i, j\}) \cup Z$, and each non-endpoint non-collider on it is not in Z or points only to nodes in the same strongly connected component of G. Such walks exist, since the 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 $\mathrm{Anc}_G(Z)$. Suppose on the contrary the existence of a collider k on μ that is not ancestor of Z. It is either ancestor of i or of j, by assumption. Without loss of generality, assume the latter. Then there is a directed path π from k to j in G that does not pass through any node of Z. Then the subwalk of μ between i and k 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 can be extended to a walk that is σ-open given Z, by replacing each collider k on it by a walk $k \to \cdots \to z \leftarrow \cdots \leftarrow k$ for some closest descendant $z \in Z$ of k (with possibly $z = k$). Hence, i and j are σ-connected given Z.

(iii) $\implies$ (iv) is trivial.

(iv) $\implies$ (i): Suppose that i and j are σ'-connected given $Z^* = \mathrm{Anc}_G(\{i, j\}) \setminus \{i, j\}$. Let π be a path between i and j that is σ'-open given Z^*. We show that π must be an inducing path. First, all colliders on π are in $\mathrm{Anc}_G(Z^*)$ and hence in $\mathrm{Anc}_G(\{i, j\})$. Second, let k be any non-endpoint non-collider on π. Then there must be a directed subwalk 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 can only point to nodes in the same strongly connected component of G. Hence, all non-endpoint non-colliders on π can only point to nodes in the same strongly connected component of G. □

We will introduce some terminology regarding the orientation of edges and of walks.

Definition 9.1.4. *Edges of the form $i \leftarrow j, i \leftrightarrow j$ are called* into *i, and similarly, edges of the form $i \to j, i \leftrightarrow j$ are called* into *j. Edges of the form $i \to j$ and $j \leftarrow i$ are called* out of *i. A walk between i and j is called* into *i if it is of the form $i \leftarrow\!\!* \cdots j$, and* out of *$i$ if it is of the form $i \to \cdots j$.*

The orientation of the outermost edges on an inducing walk (or path) contain important information.

Lemma 9.1.5. *Let $G = \langle V, E, L \rangle$ be a DMG and $i, j \in V$ be distinct. If there exists an inducing walk between i and j in G, and all inducing walks in G between i and j are out of i, then:*

1. $i \in \mathrm{Anc}_G(j)$, *and*

2. *the first node next to* i *on any inducing walk between* i *and* j *cannot be in* $\mathrm{Sc}_G(i)$.

Proof. Let μ be an inducing walk between i and j in G. It must be of the form $i \to l \cdots j$ (with possibly $l = j$). First note that l cannot be in $\mathrm{Sc}_G(i)$, because otherwise we could reverse the orientation of the first edge and obtain an inducing walk $i \leftarrow l \cdots j$ that would be into i, contradicting the assumption. If μ is a directed walk all the way to j, then clearly, $i \in \mathrm{Anc}_G(j)$. Otherwise, it must contain a collider. Let k be the collider on μ closest to i. k must be ancestor of i or j. In the latter case, clearly $i \in \mathrm{Anc}_G(j)$. In the former case, all nodes on the subwalk of μ between i and k must be in $\mathrm{Sc}_G(i)$, a contradiction. □

The same proof strategy of Proposition 9.1.2 (ii) $\implies$ (iii) can be used to show two slightly stronger statements.

Proposition 9.1.6. *Let* $G = \langle V, E, L \rangle$ *be a DMG and* $i, j \in V$ *be distinct. Then:*

(i) if there is an inducing walk in G *into* i*, then for all* $Z \subseteq V \setminus \{i, j\}$ *there exists a* Z*-*σ'*-open walk in* G *that is into* i*.*

(ii) if there is an inducing walk in G *between* i *and* j*, and all such walks are out of* i*, then for all* $Z \subseteq V \setminus \{i, j\}$ *there exists a* Z*-*σ'*-open walk of the form* $i \to k \cdots j$ *(with possibly* $k = j$*) such that* $k \notin \mathrm{Sc}_G(i)$*.*

In that latter case, we also conclude that $i \in \mathrm{Anc}_G(j)$*.*

Proof. (i) Suppose there exists an inducing walk between i and j in G that is into i. Let $Z \subseteq V \setminus \{i, j\}$. Consider all walks in G between i and j with the property that the walk is into i, all colliders on it are in $\mathrm{Anc}_G(\{i, j\}) \cup Z$, and each non-endpoint non-collider on it is not in Z or points only to nodes in the same strongly connected component of G. Such walks exist, since the 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 $\mathrm{Anc}_G(Z)$. Suppose on the contrary the existence of a collider k on μ that is not ancestor of Z. It is either ancestor of i or of j, by assumption. If it is ancestor of i, there is a directed path π from k to i in G that does not pass through any node of Z. The subwalk of μ between j and k can be concatenated with the directed path π into a walk between i and j that has the property (in particular, it is still into i), but has fewer colliders than μ: a contradiction. If it is not ancestor of i, but of j, then a similar construction can be used to arrive at a contradiction, but now using a directed path from k to j instead. Therefore, μ is σ'-open given Z, and it is into i.

(ii) Suppose there exists an inducing walk between i and j in G. Assume that all inducing walks in G between i and j are out of i. Let $Z \subseteq V \setminus \{i, j\}$. Consider all walks in G between i and j with the property that the walk is out of i, the first node next to i is not in $\mathrm{Sc}_G(i)$, all colliders on it are in $\mathrm{Anc}_G(\{i, j\}) \cup Z$, and each non-endpoint non-collider on it is not in Z or points only to nodes in the same strongly connected component of G. Such walks exist, since the inducing walk is one. Let μ be such a

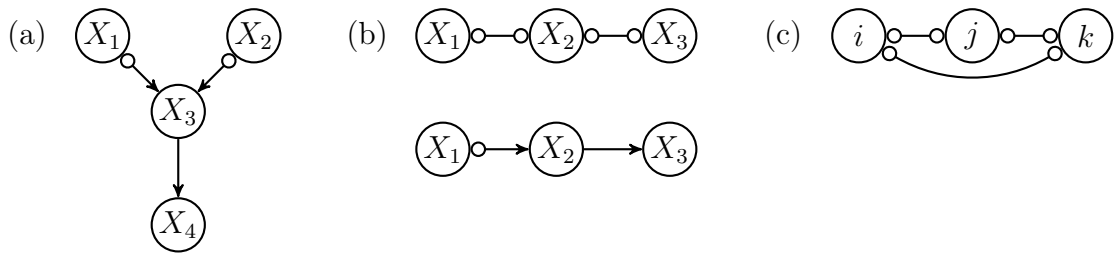


Figure 12: Example DPAGs. (a) DPAG containing the Y-structures in Figure 10. (b) Two different DPAGs that both contain all the LCD DMGs from Figure 9. (c) DPAG that contains the DMG in Figure 11(a).

walk with a minimal number of colliders. We show that all colliders on μ must be in $\text{Anc}_G(Z)$. Suppose on the contrary the existence of a collider k on μ that is not ancestor of Z. It is either ancestor of i or of j, by assumption. By Lemma 9.1.5, $i \in \text{Anc}_G(j)$. Hence, k must be ancestor of j. Therefore, there is a directed path π from k to j in G that does not pass through any node of Z. The subwalk of μ between i and k can be concatenated with the directed path π into a walk between i and j that has the property (in particular, it is still out of i, and the node next to i is unchanged), but has fewer colliders than μ: a contradiction. Therefore, μ is σ'-open given Z, and it is out of i.

The last statement follows from Lemma 9.1.5. □

9.2. Directed Partial Ancestral Graphs (DPAGs)

It is often convenient when performing causal reasoning to be able to represent a set of DMGs in a compact way. For this purpose, *partial ancestral graphs (PAGs)* have been introduced [Zha06]. Here we will assume no selection bias for simplicity, which allows us to restrict ourselves to discussing *directed* PAGs (henceforth abbreviated as DPAGs).

Directed PAGs can have multiple edge types. In addition to the three edge types we have seen before ($\to$, $\leftarrow$, $\leftrightarrow$), there are three new edge types involving a circle: $\leftarrow\!\circ$, $\circ\!\!-\!\!\circ$, $\circ\!\!\to$. Each edge $i \ast\!\!-\!\!\ast\, j$ therefore has two *edge marks*, one for 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.[28] Only 6 out of the 9 possible combinations of edge marks can occur in directed PAGs. We will make use of the "$\ast$" symbol to denote any of the three edge marks. So the notation $i \ast\!\!-\!\!\ast\, j$ includes all 6 possible edge types between i and j, whereas $i \leftarrow\!\ast\, j$ is shorthand for three possible edge types. Edges of the form $i \leftarrow j, i \leftarrow\!\circ\, j, i \leftrightarrow j$ are called *into* i, and similarly, edges of the form $i \to j, i \circ\!\!\to j, i \leftrightarrow j$ are called *into* j. Edges of the form $i \to j$ and $j \leftarrow i$ are called *out of* i.

In order to define the subclass of DPAGs, we extend the definitions of (directed) walks, (directed) paths and colliders to cover these new edge types.

[28] PAGs have more edge types than DPAGs: they can also have undirected or circle-tail edges, i.e., edges of the form $\{-, -\!\circ, \circ\!\!-\!\!\circ\}$. This means that for PAGs, all combinations of edge marks may occur.

Definition 9.2.1. *Let $G = (V, E)$ be a mixed graph with nodes V and edges E of the types $\{\rightarrow, \leftarrow, \leftarrow\!\circ, \leftrightarrow, \circ\!\!-\!\!\circ, \circ\!\!\rightarrow\}$.*[29] *Let $v, w \in V$.*

1. *If there is an edge $v \ast\!\!-\!\!\ast\, w$ between v and w (of either type), we call v and w* adjacent *in G.*

2. *A* walk between v and w *in G is a finite sequence of nodes and edges*

$$v = v_0 \ast\!\!-\!\!\ast\, v_1 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast\, v_{n-1} \ast\!\!-\!\!\ast\, v_n = w$$

   in G for some $n \geq 0$, i.e. such that for every $k = 1, \ldots, n$ we have that $v_{k-1}, v_k \in V$ and $v_{k-1} \ast\!\!-\!\!\ast\, v_k \in E$, and with $v_0 = v$ and $v_n = w$. Here, the symbol "$\ast$" can stand for any edge mark (tail, arrowhead, or circle).

3. *A walk is called a* path *if no node occurs multiple times in the walk.*

4. *A* directed walk *(path) from v to w in G is a walk (path) of the form:*

$$v = v_0 \rightarrow v_1 \rightarrow \cdots \rightarrow v_{n-1} \rightarrow v_n = w,$$

   for some $n \geq 0$.

5. *A* directed cycle *is a directed walk from $v \rightarrow \cdots \rightarrow w$, concatenated with the directed edge $w \rightarrow v$.*

6. *An* almost directed cycle *is a directed walk from $v \rightarrow \cdots \rightarrow w$, concatenated with the bidirected edge $w \leftrightarrow v$.*

7. *A triple of consecutive nodes $v_{k-1} \ast\!\!-\!\!\ast\, v_k \ast\!\!-\!\!\ast\, v_{k+1}$ on a walk is called* collider *if it is of the form $v_{k-1} \ast\!\!\rightarrow v_k \leftarrow\!\!\ast\, v_{k+1}$.*

8. *A walk $v \cdots w$ between v and w is called* inducing *if every collider on the walk is in $\mathrm{Anc}_G(\{v, w\})$, and every non-collider on the walk (except the endpoints) only has outgoing directed edges to neighboring nodes on the walk that lie in the same strongly connected component of G.*

We can now define:

Definition 9.2.2. *A mixed graph $H = \langle V, E \rangle$ with nodes V and edges E of the types $\{\rightarrow, \leftarrow, \leftarrow\!\circ, \leftrightarrow, \circ\!\!-\!\!\circ, \circ\!\!\rightarrow\}$ is called a* directed partial ancestral graph (DPAG) *if all of the following conditions hold:*

1. *Between any two distinct nodes there is at most one edge, and there are no self-cycles;*

2. *The graph contains no directed or almost directed cycles ("ancestral");*

[29] Formally, we no longer introduce separate sets to represent the edges of each type, but merge them into the single set E.

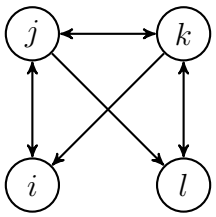


Figure 13: This mixed graph is not a valid DPAG, because it is not maximal: it has an inducing path $i \leftrightarrow j \leftrightarrow k \leftrightarrow l$ while i and l are non-adjacent.

3. *There is no inducing path between any two non-adjacent nodes ("maximal").*

We can define the skeleton of any mixed graph in the following way.

Definition 9.2.3. *Given a DPAG $H = \langle V, E\rangle$, its induced* skeleton *is the mixed graph* $\mathrm{skel}(H) = \langle V, F\rangle$ *with the same nodes, and with a bicircle edge $i \circ\!\!-\!\!\circ j$ in F if and only if $i \ast\!\!-\!\!\ast j$ 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. DPAGs are used to represent a set of directed mixed graphs as follows.

Definition 9.2.4. *Let $H = \langle V, E\rangle$ be a mixed graph with nodes V and edges E of the types $\{\rightarrow, \leftarrow, \leftarrow\!\!\circ, \leftrightarrow, \circ\!\!-\!\!\circ, \circ\!\!\rightarrow\}$, with at most one edge connecting any pair of distinct nodes. Let G be a directed mixed graph. We say that H* contains *G if all of the following hold:*

1. *G and H have the same vertex set V;*
2. *two vertices i, j are adjacent in H if and only if there is an inducing path between i, j in G;*
3. *if $i \ast\!\!\rightarrow j$ in H (i.e., $i \rightarrow j$ in H or $i \circ\!\!\rightarrow j$ in H or $i \leftrightarrow j$ in H), then $j \notin \mathrm{Anc}_G(i)$;*
4. *if $i \rightarrow j$ in H then $i \in \mathrm{Anc}_G(j)$.*

Hence, adjacencies represent inducing paths, arrowheads represent non-ancestorship, and tails represent ancestorship. Some examples are given in Figure 12. Figure 13 provides an example of a mixed graph that satisfies all conditions of a DPAG except the maximality.

We will frequently use that every mixed graph (of a certain type) that contains a DMG must be a valid DPAG.

Proposition 9.2.5. *Let $H = \langle V, E\rangle$ be a mixed graph with nodes V and edges E of the types $\{\rightarrow, \leftarrow, \leftarrow\!\!\circ, \leftrightarrow, \circ\!\!-\!\!\circ, \circ\!\!\rightarrow\}$, with at most one edge connecting any pair of distinct nodes. If H contains a directed mixed graph G, then H is a DPAG.*

Proof. "Ancestral". Suppose that H contains a directed path $i = v_1 \to \cdots \to v_n = j$. Then, each directed edge $v_k \to v_{k+1}$ along this directed path implies that $v_k \in \mathrm{Anc}_G(v_{k+1})$. By transitivity, $i \in \mathrm{Anc}_G(j)$. If H contained an edge $j \mathrel{*\!\!\to} i$, this would imply $j \notin \mathrm{Anc}_G(j)$, a contradiction. Hence such edges cannot occur. This means that no directed or almost directed cycles occur in H.

"Maximal". Suppose there were an inducing path μ in H between any two distinct nodes $u, v \in V$. Every edge $i \mathrel{*\!\!-\!\!*} j$ on μ corresponds with an inducing walk in G between i and j. By Lemma 9.2.7, these inducing walks can be chosen to be into i if the edge is $i \mathrel{\leftarrow\!\!*} j$ and also into j if the edge is $i \mathrel{*\!\!\to} j$. A collider on μ then stays a collider on the walk between u and v in G obtained by concatenating all these inducing walks. Since such colliders are in $\mathrm{Anc}_H(\{u, v\})$ by assumption, they are also in $\mathrm{Anc}_G(\{u, v\})$. All intermediate colliders on some inducing walk in G between i and j are ancestor in G of i or j, and hence, of u or v. A non-collider on μ points to a neighboring node in the same strongly-connected component of H. But H has no non-trivial strongly connected components, since we already showed that it must be acyclic. Therefore, it does not contain any non-colliders (except for the endpoints). The walk in G obtained by concatenating all inducing walks along μ is therefore actually an inducing walk in G. Hence, u, v must be adjacent in H, because H contains G. Hence all inducing paths in H are between two adjacent nodes. □

The following result shows that every DMG can be represented by a DPAG that contains it.

Proposition 9.2.6. *Let G be a directed mixed graph. There exists a DPAG* $\mathrm{DPAG}(G)$ *that contains G.*

Proof. Denote the vertex set of G as V. We will construct a mixed graph H with vertex set V as follows. Let two vertices $i, j \in V$ be adjacent in H if and only if there is an inducing path between i, j in G. In that case, orient the edge between i and j in H as follows:

$$\begin{cases} i \circ\!\!-\!\!\circ j & \text{if } i \in \mathrm{Anc}_G(j) \text{ and } j \in \mathrm{Anc}_G(i), \\ i \to j & \text{if } i \in \mathrm{Anc}_G(j) \text{ and } j \notin \mathrm{Anc}_G(i), \\ i \leftarrow j & \text{if } i \notin \mathrm{Anc}_G(j) \text{ and } j \in \mathrm{Anc}_G(i), \\ i \leftrightarrow j & \text{if } i \notin \mathrm{Anc}_G(j) \text{ and } j \notin \mathrm{Anc}_G(i). \end{cases}$$

It is obvious by construction that H contains G. It is a valid DPAG by Proposition 9.2.5. □

If G is acyclic, the DPAG constructed in this way contains no circle edge marks (and is called the directed maximal ancestral graph (DMAG) induced by G in the literature).

The following lemma shows that the orientation of edges in a DPAG that contains a DMG contains information on the orientation of corresponding inducing paths in the DMG.

Lemma 9.2.7. *Let H be a DPAG that contains DMG G. If $k \mathrel{*\!\!\to} i$ in H, then there exists an inducing walk in G between k and i that is into i. If $k \leftrightarrow i$ in H, then there exists an inducing walk in G between k and i that is both into k and into i.*

Proof. If $k \ast\!\!\rightarrow i$ in H, then there exists an inducing walk between k and i in G because k and i are adjacent in H and H contains G. If this inducing walk were out of i, it would be of the form $k \ldots \ast\!\!\rightarrow u_n \leftarrow u_{n-1} \leftarrow \cdots \leftarrow u_1 \leftarrow i$, where u_n is the first collider on the walk that one encounters when following the directed edges out of i (note that it cannot be a directed walk from i to k because then $i \notin \mathrm{Anc}_G(k)$, contradicting the orientation $k \ast\!\!\rightarrow i$ in H). u_n must be ancestor of i or k in G, and it cannot be ancestor of k (because then i would be ancestor of k, contradicting the orientation $k \ast\!\!\rightarrow i$ in H), hence it must be ancestor of i. Therefore, all nodes $i, u_1, \ldots, u_n$ lie in the same strongly connected component of G. Thus there exists a walk $k \ldots \ast\!\!\rightarrow u_n \rightarrow \cdots \rightarrow i$ in G where we replaced the subwalk $u_n \leftarrow u_{n-1} \leftarrow \cdots \leftarrow i$ by a directed path from u_n to i. It is clear that this is an inducing walk in G between k and i that is into i.

If $k \leftrightarrow i$ in H, then by similar reasoning, we obtain an inducing walk in G between k and i that is into k as well as into i. □

9.3. Unshielded triples

One of the key steps in the FCI algorithm is the orientation of "unshielded triples".

Definition 9.3.1. *Let H be a mixed graph. A triple of distinct nodes $\langle i, j, k\rangle$ in H is called an* unshielded triple *if $i \ast\!\!-\!\!\ast j$ in H, and $j \ast\!\!-\!\!\ast k$ in H, but i and k are not adjacent in H.*

The following proposition will later be used to "orient" the edges in unshielded triples in a DPAG containing a DMG.

Proposition 9.3.2. *Let H be a DPAG that contains a DMG G with vertex set V. If $\langle i, j, k\rangle$ form an unshielded triple in H, then either*

(i) $j \notin Z$ for each $Z \subseteq V \setminus \{i, k\}$ that σ-separates i from k in G, and $j \notin \mathrm{Anc}_G(\{i, k\})$, or

(ii) $j \in Z$ for each $Z \subseteq V \setminus \{i, k\}$ that σ-separates i from k in G, and $j \in \mathrm{Anc}_G(\{i, k\})$.

Proof. Because H contains G and $\langle i, j, k\rangle$ is an unshielded triple in H, there must exist an inducing walk in G between i and j, and one between j and k, but there exists no inducing walk in G between i and k.

Suppose there are inducing walks of the form $i \cdots \ast\!\!\rightarrow j$ and $j \leftarrow\!\!\ast \cdots k$ in G, i.e., both into j. Let $Z \subseteq V \setminus \{i, k\}$ be such that $i \perp_G^{\sigma} k \,|\, Z$. By Proposition 9.1.6, we can find a Z-σ'-open walk in G between i and j that is into i, and one between j and k that is into j. By concatenating the two walks, we obtain a walk between i and k in G that has j as a collider, and that is Z-σ'-open if and only if $j \in \mathrm{Anc}_G(Z)$. Since we assumed that Z σ-separates i from k in G, we get $j \notin \mathrm{Anc}_G(Z)$. In particular, this implies that $j \notin Z$. Because there is no inducing walk between i and k in G, Proposition 9.1.2 tells us that $i \perp_G^{\sigma} k \,|\, \mathrm{Anc}_G(\{i, k\}) \setminus \{i, k\}$. For the special case $Z = \mathrm{Anc}_G(\{i, k\}) \setminus \{i, k\}$ we then conclude that $j \notin \mathrm{Anc}_G(\{i, k\})$.

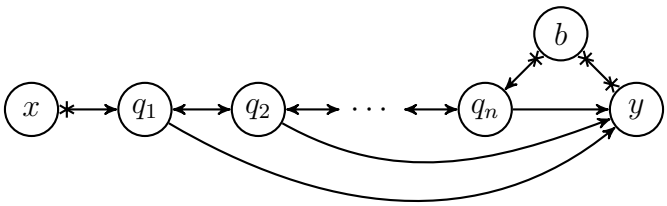


Figure 14: Discriminating path $\langle x, q_1, q_2, \ldots, q_n, b, y\rangle$ for b between x and y.

Alternatively, suppose that there are no inducing walks in G between i and j that are into j, or no inducing walks in G between j and k that are into j. Without loss of generality, assume the former. Then all inducing walks in G between i and j must be out of j. By Proposition 9.1.6, then $j \in \mathrm{Anc}_G(i)$. Let $Z \subseteq V \setminus \{i, k\}$ be such that $i \perp^{\sigma}_G k \,|\, Z$. By Proposition 9.1.6, we can find a Z-σ'-open walk in G between i and j that is out of j and on which j points to a node in another strongly connected component of G. By Proposition 9.1.2 we can also find a Z-σ'-open walk in G between j and k. By concatenating the two walks, we obtain a walk in G between i and k that has j as a non-endpoint non-collider, and that is Z-σ'-open if and only if $j \notin Z$ or j only points to nodes in the same strongly connected component of G. Since we assumed that Z σ-separates i from k in G, we conclude that $j \in Z$. □

Note that in the first case, we can orient the edges as $i \mathrel{*\!\!\rightarrow} j \mathrel{\leftarrow\!\!*} k$ (if they were not already oriented in this way) to obtain a DPAG $\tilde{H}$ that also contains G. In the second case, we cannot orient the edges, since we don't know whether $j \in \mathrm{Anc}_G(i)$ or $j \in \mathrm{Anc}_G(k)$ or both.

9.4. Discriminating paths

Another step in FCI is related to the notion of "discriminating paths". This can be considered as an extension of the notion of unshielded triple.

Definition 9.4.1. *A path $\pi = \langle x, q_1, \ldots, q_n, b, y\rangle$ (with $n \geq 1$) in a DPAG H is a* discriminating path for b *if:*

(i) x is not adjacent to y in H, and

(ii) every node q_i $(1 \leq i \leq n)$ is a collider on π and a parent of y in H.

Figure 14 illustrates this notion.

Remark 9.4.2. *It is instructive to think about a discriminating path rather as a certain*

collection of paths:

$$
\begin{aligned}
&x \mathrel{*\!\!\rightarrow} q_1 \rightarrow y \\
&x \mathrel{*\!\!\rightarrow} q_1 \leftrightarrow q_2 \rightarrow y \\
&\vdots \\
&x \mathrel{*\!\!\rightarrow} q_1 \leftrightarrow q_2 \leftrightarrow \cdots \leftrightarrow q_n \rightarrow y \\
&x \mathrel{*\!\!\rightarrow} q_1 \leftrightarrow q_2 \leftrightarrow \cdots \leftrightarrow q_n \mathrel{\leftarrow\!\!*} b \mathrel{*\!\!-\!\!*} y
\end{aligned}
$$

with the additional requirement that x and y are not adjacent.

Remark 9.4.3. *Note that the discriminating must always be into y if H contains a DMG G. Indeed, if y were in $\mathrm{Anc}_G(b)$, then $q_n \in \mathrm{Anc}_G(b)$ because $q_n \in \mathrm{Anc}_G(y)$, contradicting the arrowhead at q_n in $q_n \mathrel{\leftarrow\!\!*} b$.*

The following quintessential property of discriminating paths is analogous to that of unshielded triples.

Proposition 9.4.4. *Let H be a DPAG that contains a DMG G with vertex set V. If $\langle i, q_1, \ldots, q_n, j, k\rangle$ is a discriminating path in H for j between i and k, then either*

(i) $j \notin Z$ for each $Z \subseteq V \backslash \{i, k\}$ that σ-separates i from k in G, and $j \notin \mathrm{Anc}_G(\{q_n, k\})$, or

(ii) $j \in Z$ for each $Z \subseteq V \setminus \{i, k\}$ that σ-separates i from k in G, and $j \in \mathrm{Anc}_G(k)$.

In both cases, $k \notin \mathrm{Anc}_G(j)$.

Proof. Because the DPAG H contains DMG G, every edge $i \mathrel{*\!\!-\!\!*} j$ in the discriminating path in H corresponds to an inducing walk in G. With Lemma 9.2.7, we conclude that these inducing walks can be chosen to be into i if $i \mathrel{\leftarrow\!\!*} j$ in H and into j if $i \mathrel{*\!\!\rightarrow} j$ in H. Furthermore, each $q_i \in \mathrm{Anc}_G(y)$ for $i = 1, \ldots, n$. For all $i = 1, \ldots, n$, there cannot exist inducing walks in G between q_i and y that are into q_i. Indeed, if there were such a walk, one could concatenate the inducing walks in G between $x, q_1, \ldots, q_i$ and y into one inducing walk in G between x and y, contradicting their non-adjacency in H. Therefore, all inducing walks in G between q_i and y must be out of q_i, for all $i = 1, \ldots, n$.

Suppose that $Z \subseteq V \setminus \{x, y\}$ σ-separates x and y in G. By Proposition 9.1.6, We can find Z-σ'-open walks between i and j replacing each adjacent $i \mathrel{*\!\!-\!\!*} j$ in the paths above, which are into i if the edge is into i and into j if the edge is into j. For the edges $q_i \mathrel{*\!\!-\!\!*} b$ we can find a Z-σ'-open walk between q_i and y that is out of q_i. For each of the paths in Remark 9.4.2, consider the corresponding concatenation of such Z-σ'-open walks in G. Since x and y are non-adjacent in H, and H contains G, these concatenated walks must be Z-σ'-closed. For the first path in Remark 9.4.2, this implies that $q_1 \in Z$. For the second path, this then means that $q_2 \in Z$. Repeating this reasoning, we conclude that all $q_i \in Z$ for $i = 1, \ldots, n$.

For the last path that includes b, we now reason similarly as for the unshielded triple in Proposition 9.3.2. Suppose there exists an inducing walk in G between q_n and b that

is into b, and an inducing walk in G between b and y that is into b. Let $Z \subseteq V \setminus \{x, y\}$ be such that $x \perp^{\sigma}_{G} y \,|\, Z$. Concatenating all corresponding Z-σ'-open walks in G between the all subsequent vertices of the discriminating path $x, q_1, \ldots, q_n, b, y$, we get a walk in G between x and y on which b is a collider. The only way in which this walk could be Z-σ'-closed is if $b \notin \mathrm{Anc}_G(Z)$. Hence we need $b \notin Z$. In particular, taking $Z = \mathrm{Anc}_G(\{x, y\}) \setminus \{x, y\}$ (which must σ-separate x from y in G because we assumed that x and y are non-adjacent in H), we conclude that $b \notin \mathrm{Anc}_G(\{x, y\})$. In particular, this means that $b \notin \mathrm{Anc}_G(y)$. Since $q_n \in \mathrm{Anc}_G(y)$, this also implies that $b \notin \mathrm{Anc}_G(q_n)$.

Suppose now instead that all inducing walks in G between b and q_n are out of b. Then $b \in \mathrm{Anc}_G(q_n)$ by Proposition 9.1.6. Since $q_n \in \mathrm{Anc}_G(y)$, we get $b \in \mathrm{Anc}_G(y)$. Let $Z \subseteq V \setminus \{x, y\}$ be such that $x \perp^{\sigma}_{G} y \,|\, Z$. In this case, b becomes a noncollider on the concatenated walk between x and y, and in order to Z-σ'-close the walk, we need that $b \in Z$.

The last case to consider is that all inducing walks between b and y are out of b. Then $b \in \mathrm{Anc}_G(y)$ by Proposition 9.1.6. Let $Z \subseteq V \setminus \{x, y\}$ be such that $x \perp^{\sigma}_{G} y \,|\, Z$. b again becomes a noncollider on the concatenated walk between x and y, and in order to Z-σ'-close the walk, we need that $b \in Z$. □

Note that in the first case, we can orient $q_n \leftrightarrow b \leftrightarrow y$ (as far as the edge marks were not already oriented in this way) to obtain a DPAG $\tilde{H}$ that also contains G. In the second case, we can orient $b \rightarrow y$ (as far as the edge marks were not already oriented in this way) to obtain a DPAG $\tilde{H}$ that also contains G.

9.5. Independence models and Markov equivalence

Definition 9.5.1. *For a DMG $G = \langle V, E, L \rangle$, define its* d-independence model *to be*

$$\mathrm{IM}_d(G) := \{\langle A, B, C \rangle : A, B, C \subseteq V, A \overset{d}{\underset{G}{\perp}} B \mid C\},$$

i.e., the set of all d-separations entailed by the graph. Define its σ-independence model *to be*

$$\mathrm{IM}_\sigma(G) := \{\langle A, B, C \rangle : A, B, C \subseteq V, A \overset{\sigma}{\underset{G}{\perp}} B \mid C\},$$

i.e., the set of all σ-separations entailed by the graph.

In general, given an index set V, we call a subset of $\mathcal{P}(V)^3 = \{\langle A, B, C \rangle : A, B, C \subseteq V\}$ an *independence model over* V.[30] If $\langle A, B, C \rangle$ in an independence model, we also say that C separates A from B. Both $\mathrm{IM}_d(G)$ and $\mathrm{IM}_\sigma(G)$ are independence models over V, the vertices of G. For ADMGs, σ-separation is equivalent to d-separation, and hence, if G is acyclic, then $\mathrm{IM}_d(G) = \mathrm{IM}_\sigma(G)$.

The input to the FCI algorithm will consist of an independence model over V, and its output will consist of a mixed graph with vertices V. If the input of FCI is the

[30] Here, $\mathcal{P}(V)$ denotes the power set of V, i.e., the set of all subsets of V, rather than the space of probability distributions on V.

independence model of a DMG, then the output of FCI will be a DPAG that represents the "Markov equivalence class" of the DMG, as we will see later.

Definition 9.5.2. *We call two DMGs G_1 and G_2* σ-Markov equivalent *if* $\mathrm{IM}_\sigma(G_1) = \mathrm{IM}_\sigma(G_2)$*, and* d-Markov equivalent *if* $\mathrm{IM}_d(G_1) = \mathrm{IM}_d(G_2)$.

9.6. FCI Algorithm

We need one more definition before we can state the FCI algorithm.

Definition 9.6.1. *A path $v_0 \ast\!\!-\!\!\ast\, v_1 \ast\!\!-\!\!\ast \cdots \ast\!\!-\!\!\ast\, v_n$ between nodes v_0 and v_n in a DPAG H is called a* possibly directed path from v_0 to v_n *if for each $i = 1, \ldots, n$, the edge $v_{i-1} \ast\!\!-\!\!\ast\, v_i$ is not into v_{i-1} (i.e., is of the form $v_{i-1} \circ\!\!-\!\!\circ v_i$, $v_{i-1} \circ\!\!\rightarrow v_i$, or $v_{i-1} \rightarrow v_i$). The path is called* uncovered *if every subsequent triple $\langle v_{i-1}, v_i, v_{i+1}\rangle$ is unshielded, i.e., v_{i-1} and v_{i+1} are not adjacent in H for $i = 1, \ldots, n-1$.*

We are now ready to describe a causal inference algorithm that is closely related to FCI. Its input is an independence model over an index set V. Its output is a mixed graph with vertex set V. It starts with a *skeleton phase* that is aimed at deducing the adjacencies between the nodes, and to find sets that separate two nodes. Then, it runs various *orientation rules* that iteratively orients edge marks into tails and arrowheads.

1. INPUT: Vertex set V, independence model I over index set V.
2. Initialize H to be the complete graph on V with an edge $\circ\!\!-\!\!\circ$ between each pair of distinct nodes.
3. For each unordered pair $i \neq j$ of nodes in H:
   search for a subset $S \subseteq V \setminus \{i, j\}$ such that $\langle i, j, S\rangle \in I$; if such a set S is found then remove the edge $i \circ\!\!-\!\!\circ j$ from H and record S in $\mathrm{sepset}(\{i, j\})$.
4. Orient unshielded triples in H:
   - $\mathcal{R}0$ for each unshielded triple $\langle i, j, k\rangle$ in H, orient it as a collider $i \ast\!\!\rightarrow j \leftarrow\!\!\ast\, k$ if and only if j is not in $\mathrm{sepset}(\{i, k\})$.
5. Perform the following orientation rules until none of them applies:
   - $\mathcal{R}1$ If $i \ast\!\!\rightarrow j \circ\!\!-\!\!\ast\, k$ in H, and i and k are not adjacent in H, orient the triple as $i \ast\!\!\rightarrow j \rightarrow k$.
   - $\mathcal{R}2$ 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 $i \ast\!\!\rightarrow j \leftarrow\!\!\ast\, k$ in H, and $i \ast\!\!-\!\!\circ l \circ\!\!-\!\!\ast\, k$ in H, and $l \ast\!\!-\!\!\circ j$ in H, and i and k are not adjacent in H, then orient $l \ast\!\!\rightarrow j$.
   - $\mathcal{R}4$ If $\langle i, q_1, \ldots, q_n, j, k\rangle$ is a discriminating path in H for j, and if $j \circ\!\!-\!\!\ast\, k$ in H, then orient $j \rightarrow k$ if $j \in \mathrm{sepset}(\{i, k\})$ and orient $q_n \leftrightarrow j \leftrightarrow k$ if $j \notin \mathrm{sepset}(\{i, k\})$.

6. Perform the following orientation rules until none of them applies:

   $\mathcal{R}8$ If $i \to j \to k$ in H, and $i \circ\!\!\to k$ in H, then orient $i \to k$.

   $\mathcal{R}9$ If $i \circ\!\!\to k$, and $\pi = \langle i, j, \ldots, k \rangle$ is an uncovered possibly directed path in H from i to k such that j and k are not adjacent in H, then orient $i \to k$.

   $\mathcal{R}10$ Suppose that $i \circ\!\!\to k$ in H, $j \to k \leftarrow l$ in H, π_1 is a uncovered possibly directed path in H from i to j, and π_2 is a u.p.d. path in H from i to l. Let u_1 be the vertex adjacent to i on π_1 (possibly $u_1 = j$) and u_2 the vertex 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 $i \to k$.

7. OUTPUT the mixed graph H.

We have here omitted orientation rules $\mathcal{R}5$–$\mathcal{R}7$ since these are only necessary if selection bias is not ruled out a priori. Those rules can yield edges of the form $-\!\!\circ, \circ\!\!-$ and $-$ that are not valid in DPAGs, and require the more general formalism of PAGs.

We did not specify yet how the search for a separating set is performed in the so-called skeleton phase. A brute-force search over all possible subsets of $V \setminus \{i, j\}$ is not only computationally extremely expensive for all but the largest cardinalities of V, but also statistically not very reliable. There have been three proposals of improved search techniques, but the details of these are somewhat too involved to reproduce here. The original proposal motivated the (somewhat optimistic) adjective "Fast" in the name of the FCI algorithm and involves the so-called "Possible-D-Sep" sets. Later, an alternative search strategy was proposed that can be considerably faster in practice than FCI, without sacrificing completeness, and 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. Another variant, the "Really Fast Causal Inference" algorithm has also been proposed, which adapts both the skeleton phase and the orientation rules, and sacrifices some completeness but remains sound.

9.7. Soundness and completeness of FCI

Theorem 9.7.1. *The FCI algorithm is* sound, *i.e., if its input consists of $I_\sigma(G)$ for a DMG G with vertex set V, then its output will be a valid DPAG H that contains G.*

Proof. Since the details of the skeleton phase are omitted, we will assume that the implementation of that step is sound, or a brute-force search is done. We therefore assume that after step 3 of the FCI algorithm, the mixed graph H has vertex set V and has an edge between any pair of distinct nodes $i, j \in V$ if and only if there is no set $S \subseteq V \setminus \{i, j\}$ such that $i \perp^\sigma_G j \,|\, S$. Furthermore, if there is no edge in H between a pair of distinct nodes $i, j \in V$, then $i \perp^\sigma_G j \,|\, \mathrm{sepset}(\{i, j\})$. By Proposition 9.1.2, this implies that two distinct nodes $i, j \in V$ are adjacent in H at this stage if and only if there is an inducing walk in G between i and j. By construction, the only edge type occurring at this stage in H is $\circ\!\!-\!\!\circ$, and hence we conclude that H is a valid DPAG that contains G.

Every application of rule $\mathcal{R}0$ in step 4 adapts H by turning certain circle edge marks into arrowheads. This does not invalidate the validity of the DPAG H, and it follows from Proposition 9.3.2 that after each of these orientations, H still contains G.

The proof proceeds by induction. We just show for each of the orientation rules that under the assumption that the current H is a valid DPAG that contains G, applying the rule yields an updated H that is still a valid DPAG that contains G. In the following, we will always assume that the antecedent of the rule holds for a mixed graph H that is a valid DPAG that contains DMG G.

$\mathcal{R}1$ "If $i \ast\!\!\rightarrow j \circ\!\!-\!\!\ast\, k$ in H, and i and k are not adjacent in H, orient the triple as $i \ast\!\!\rightarrow j \rightarrow k$."
We have to show that $j \in \mathrm{Anc}_G(k)$. Since the triple $\langle i, j, k\rangle$ is an unshielded triple in H, but has not been oriented as a collider by $\mathcal{R}0$, we conclude that $j \in \mathrm{sepset}(\{i, k\})$. By Proposition 9.3.2, $j \in \mathrm{Anc}_G(\{i, k\})$. Since H contains G and $i \ast\!\!\rightarrow j$ in H, $j \notin \mathrm{Anc}_G(i)$. Therefore, $j \in \mathrm{Anc}_G(k)$. After orienting $j \rightarrow k$ in H, the resulting mixed graph H still contains G.

$\mathcal{R}2$ "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 $i \ast\!\!\rightarrow k$."
By the antecedent of the rule, and since H contains G, we have $k \notin \mathrm{Anc}_G(j)$. In case $i \rightarrow j \ast\!\!\rightarrow k$, we have $i \in \mathrm{Anc}_G(j)$, so if k were in $\mathrm{Anc}_G(i)$, it would follow that $k \in \mathrm{Anc}_G(j)$, a contradiction. In case $i \ast\!\!\rightarrow j \rightarrow k$, we have on one hand $j \notin \mathrm{Anc}_G(i)$, and on the other $j \in \mathrm{Anc}_G(k)$, so if k were in $\mathrm{Anc}_G(i)$, it would follow that $j \in \mathrm{Anc}_G(i)$, contradicting $j \notin \mathrm{Anc}_G(i)$. Hence, in both cases, we must have $k \in \mathrm{Anc}_G(i)$. After orienting $i \ast\!\!\rightarrow k$ in H, the resulting mixed graph still contains G.

$\mathcal{R}3$ "If $i \ast\!\!\rightarrow j \leftarrow\!\!\ast\, k$ in H, and $i \ast\!\!-\!\!\circ\, l \circ\!\!-\!\!\ast\, k$ in H, and $l \ast\!\!-\!\!\circ\, j$ in H, and i and k are not adjacent in H, then orient $l \ast\!\!\rightarrow j$."
Since $\langle i, l, k\rangle$ is an unshielded triple in H that was not oriented as a collider by $\mathcal{R}0$, we must have that $l \in \mathrm{Anc}_G(\{i, k\})$ by Proposition 9.3.2. Assume, for the sake of contradiction, that $j \in \mathrm{Anc}_G(l)$. Then $j \in \mathrm{Anc}_G(\{i, k\})$. This contradicts that $j \notin \mathrm{Anc}_G(\{i, k\})$ from $i \ast\!\!\rightarrow j \leftarrow\!\!\ast\, k$ in H. Hence, $j \notin \mathrm{Anc}_G(l)$. After orienting $l \ast\!\!\rightarrow j$ in H, the resulting mixed graph still contains G.

$\mathcal{R}4$ "If $\langle i, q_1, \ldots, q_n, j, k\rangle$ is a discriminating path in H for j, and if $j \circ\!\!-\!\!\ast\, k$ in H, then orient $j \rightarrow k$ if $j \in \mathrm{sepset}(\{i, k\})$ and orient $q_n \leftrightarrow j \leftrightarrow k$ if $j \notin \mathrm{sepset}(\{i, k\})$."
It follows immediately from Proposition **??** that applying this rule yields an updated mixed graph that still contains G.

$\mathcal{R}8$ "If $i \rightarrow j \rightarrow k$ in H, and $i \circ\!\!\rightarrow k$ in H, then orient $i \rightarrow k$."
Clearly, $i \in \mathrm{Anc}_G(j)$ and $j \in \mathrm{Anc}_G(k)$ implies $i \in \mathrm{Anc}_G(k)$. Thus applying this rule yields an updated mixed graph that still contains G.

$\mathcal{R}9$ "If $i \circ\!\!\rightarrow k$, and $\pi = \langle i, j, \ldots, k\rangle$ is an uncovered possibly directed path in H from i to k such that j and k are not adjacent in H, then orient $i \rightarrow k$."
Suppose $i \notin \mathrm{Anc}_G(k)$. The unshielded triple $\langle j, i, k\rangle$ was not oriented as a collider

by $\mathcal{R}0$, and therefore $i \in \mathrm{Anc}_G(j)$. We can repeat such reasoning for each subsequent unshielded triple along the uncovered possibly directed path between i and k. Overall, by transitivity this implies that $i \in \mathrm{Anc}_G(k)$. This, however, contradicts the assumption $i \notin \mathrm{Anc}_G(k)$. Therefore, the only possibility is $i \in \mathrm{Anc}_G(k)$. Applying the rule therefore yields an updated mixed graph that still contains 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 a u.p.d. path in H from i to l. Let u_1 be the vertex adjacent to i on π_1 (possibly $u_1 = j$) and u_2 the vertex 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 $i \rightarrow k$." The unshielded triple $\langle u_1, i, u_2 \rangle$ that was not oriented by rule $\mathcal{R}0$ implies that $i \in \mathrm{Anc}_G(u_1)$ or $i \in \mathrm{Anc}_G(u_2)$. If $i \in \mathrm{Anc}_G(u_1)$, similar reasoning for each subsequent unshielded triple along the uncovered possibly directed path π_1 leads us to conclude by transitivity that $i \in \mathrm{Anc}_G(j)$, and since $j \in \mathrm{Anc}_G(k)$, also $i \in \mathrm{Anc}_G(k)$. In case $i \in \mathrm{Anc}_G(u_2)$, analogous reasoning for π_2 leads to the same conclusion, $i \in \mathrm{Anc}_G(k)$. In both cases, applying the rule yields an updated mixed graph that still contains G.

Since each of these orientation rules leaves the skeleton (adjacencies) of H invariant, and each orientation of an edge mark leads to a valid edge, by Proposition 9.2.5, and after each orientation rule, H still contains G, H remains a valid DPAG throughout the orientation phase. □

For acyclic G, 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 $\mathrm{IM}_\sigma(G)$ will be oriented. Using results on the characterization of Markov equivalence classes of DMAGs [ARSZ05], it can be additionally shown that the DPAG output by FCI represents the Markov equivalence class of G in case G is acyclic. These completeness results are rather nontrivial and require very long proofs, and we will therefore not provide these here. By employing acyclifications, these results could easily be extended to cyclic G [MC20]. We will only state the completeness results here, and refer the interested reader to the original papers for the proofs.

Consider FCI as a mapping FCI that maps an independence model (on V) to a mixed graph (with vertex set V). It maps the independence model of a DMG G to the DPAG $\mathrm{FCI}(\mathrm{IM}_\sigma(G))$. For two DMGs G_1, G_2 with the same vertex set V, we will write $G_1 \overset{\sigma}{\sim} G_2$ if $\mathrm{IM}_\sigma(G_1) = \mathrm{IM}_\sigma(G_2)$, i.e., if G_1 and G_2 are σ-Markov equivalent.

Theorem 9.7.2. *The FCI algorithm is:*

(i) arrowhead complete*: for all DMGs G, if $i \notin \mathrm{Anc}_{\tilde{G}}(j)$ for all DMGs $\tilde{G} \overset{\sigma}{\sim} G$, then there is an arrowhead $i \leftarrow\!\!\ast\ j$ in* $\mathrm{FCI}(\mathrm{IM}_\sigma(G))$*;*

(ii) tail complete*: for all DMGs G, if $i \in \mathrm{Anc}_{\tilde{G}}(j)$ for all DMGs $\tilde{G} \overset{\sigma}{\sim} G$, then there is a tail $i \rightarrow j$ in* $\mathrm{FCI}(\mathrm{IM}_\sigma(G))$*;*

(iii) Markov complete*: for all DMGs G_1 and G_2, $G_1 \overset{\sigma}{\sim} G_2$ if and only if* $\mathrm{FCI}(\mathrm{IM}_\sigma(G_1)) = \mathrm{FCI}(\mathrm{IM}_\sigma(G_2))$.

Proof. We refer the reader to [MC20]. □

Arrowhead and tail completeness express that the DPAG output by FCI is maximally oriented: any arrowhead or tail that could possibly be deduced from $\mathrm{IM}_\sigma(G)$, will be oriented. The soundness and Markov completeness properties together imply that the DPAG $\mathrm{FCI}(\mathrm{IM}_\sigma(\mathcal{G}))$ output by FCI, when given as input the σ-independence model of a DMG $\mathcal{G}$, represents the σ-Markov equivalence class of $\mathcal{G}$. In other words, FCI provides a *graphical characterization* of the σ-Markov equivalence class of a DMG.

While the soundness of FCI allows us to directly read off some (non-)ancestral relations from the DPAG output by FCI, this is not all causal information that is identifiable from the σ-Markov equivalence class. Indirectly, one can also read off additional (non-)ancestral relations, direct causal relations, the absence of confounding, and the absence of cycles. For details, see [MC20].

In practice, of course we often do not know the independence model $\mathrm{IM}_\sigma(G)$, and have to resort to testing conditional independences in finite samples. The "oracle" soundness and completeness results formulated above, combined with asymptotically consistent conditional independence tests, leads (assuming σ-faithfulness) directly to the asymptotic consistency of the FCI algorithm. We formulate this as the following corollary.

Corollary 9.7.3. *Let $\mathcal{M}$ be a simple SCM with observed variables $O \subseteq V \cup W$, and without exogenous input variables (i.e., $J = \emptyset$). Assume that $\mathcal{M}$ is σ-faithful w.r.t. O. Denote the observable causal graph as $G := G^O(\mathcal{M})$. When using asymptotically consistent conditional independence tests on i.i.d. samples of the observational distribution $P_{\mathcal{M}}(X_O)$, FCI provides an asymptotically consistent estimate $\hat{H}$ of the DPAG $\mathrm{FCI}(\mathrm{IM}_\sigma(G))$ that represents the σ-Markov equivalence class of G. From the estimated DPAG $\hat{H}$, we can obtain consistent estimates for:*

(i) the absence/presence of (possibly indirect) causal relations according to $\mathcal{M}$;

(ii) the absence of confounding according to $\mathcal{M}$;

(iii) the absence/presence of direct causal relations according to $\mathcal{M}$;

(iv) the absence of causal cycles according to $\mathcal{M}$.

Proof. We refer the reader to [MC20]. □

Note that this corollary also applies to L-CBNs, as these can be considered to be special cases of simple SCMs.

9.8. Skeleton phase

We will now take a closer look at the skeleton search phase of the FCI algorithm.

Definition 9.8.1. *Let $G = \langle V, E, L \rangle$ be directed mixed graph (DMG). For $i, j \in V$, define $\mathrm{DSEP}_G(i, j)$ to be the set of nodes $v \in V$ such that $v \neq i$ and there is a walk π in G between i and v such that every node on π is in $\mathrm{Anc}_G(\{i, j\})$, and every non-endpoint*

non-collider on π only has outgoing directed edges to neighboring nodes on π in the same strongly connected component of G.

Note that if $j \in \mathrm{DSEP}_G(i, j)$, there is an inducing walk in G between i, j.

Proposition 9.8.2. *Let $G = \langle V, E, L \rangle$ be directed mixed graph (DMG). Let $i, j \in V$ be distinct. If there is no inducing walk between i, j in G, then i and j are σ-separated in G given $\mathrm{DSEP}_G(i, j)$, where $\mathrm{DSEP}_G(i, j) \cap \{i, j\} = \emptyset$.*

Proof. Suppose there is no inducing walk between i, j in G. Hence, $j \notin \mathrm{DSEP}_G(i, j)$. By definition, $i \notin \mathrm{DSEP}_G(i, j)$. We prove the σ-separation by contradiction. Suppose there exists a walk π in G between i and j that is σ-open given $\mathrm{DSEP}_G(i, j)$. We may assume that π contains i and j both only once.

We first show that every node on π must be in $\mathrm{Anc}_G(\{i, j\})$. This is trivial for the endpoints i, j. Every collider on π must be in $\mathrm{DSEP}_G(i, j) \subseteq \mathrm{Anc}_G(\{i, j\})$. For every non-endpoint non-collider on π, there must exist a directed subwalk of π starting at that node and ending either at a collider on π or at an end node of π, and hence it must also be in $\mathrm{Anc}_G(\{i, j\})$.

Number the nodes on π subsequently as $i = v_0, v_1, \ldots, v_n = j$, with $n > 1$. We will show that for all $k = 1, \ldots, n$, the subwalk from v_0 to v_k on π has the property that all nodes on it are in $\mathrm{Anc}_G(\{i, j\})$ and every non-endpoint non-collider only has outgoing directed edges to neighboring nodes in the same strongly connected component of G (and hence $v_1, \ldots, v_k$ are in $\mathrm{DSEP}_G(i, j)$). It is clear that this property holds for $k = 1$. Suppose it holds for $k < n$. Then $v_1, \ldots, v_k$ are all in $\mathrm{DSEP}_G(i, j)$. We will show the property holds for $k + 1$. Consider the subwalk π' of π from v_0 to v_{k+1}. If v_k is a collider on π', the property obviously holds by the induction hypothesis. If v_k is a non-collider on π', it only points to neighboring nodes on π in the same strongly connected component of G because π is $\mathrm{DSEP}_G(i, j)$-σ-open and $v_k \in \mathrm{DSEP}_G(i, j)$ is a non-endpoint non-collider on π'. In both cases, the property holds.

Hence all nodes on π are in $\mathrm{DSEP}_G(i, j)$ except for i itself. This means that in particular $j \in \mathrm{DSEP}_G(i, j)$, a contradiction. □

We say that three nodes in a graph form a triangle if each pair of nodes in the triple is adjacent. In practice, one does not know the set $\mathrm{DSEP}_G(i, j)$ if G is unknown. One can, however, easily obtain a "bound" on this set by identifying a superset.

Definition 9.8.3. *Let $H = \langle V, E \rangle$ be a mixed graph with nodes V and edges E of the types $\{\rightarrow, \leftarrow, \leftarrow\!\circ, \leftrightarrow, \circ\!\!-\!\!\circ, \circ\!\rightarrow\}$, with at most one edge connecting any pair of distinct nodes. For $i, j \in V$ distinct, we define $\mathrm{PossibleDSep}_H(i, j) \subseteq V$ to consist of those nodes $v \in V$ such that $v \neq i$ and there is a path between i and v in H such that for every subsequent triple $a \ast\!\!-\!\!\ast\, b \ast\!\!-\!\!\ast\, c$ on the path, either b is a collider in H, or a, b, c form a triangle in H.*

We can now show that

Proposition 9.8.4. *Let H be a mixed graph constructed by the PC skeleton phase followed by applying orientation rule $\mathcal{R}0$ on it, with as input $\mathrm{IM}_\sigma(G)$ for some DMG G with vertex set V. (In other words, in the first steps of FCI). For $i, j \in V$ distinct, $\mathrm{DSEP}_G(i, j) \subseteq \mathrm{PossibleDSep}_H(i, j)$.*

Proof. Note that the skeleton of H is a supergraph of the skeleton of $\mathrm{DPAG}(G)$. If $v \in \mathrm{DSEP}_G(i, j)$ there is a path π in G between i and v. There is a corresponding path π' in H between i and v that consists of the same sequence of nodes (but may have different edges between the nodes in general). Consider a subsequent triple $a \ast\!\!-\!\!\ast\, b \ast\!\!-\!\!\ast\, c$ on π. Suppose b is a collider on π. Also in H, a must be adjacent to b, and b to c. If a and c are not adjacent in H, then there must be a separating set, and hence there cannot be an inducing path between a and c in G. Therefore, rule $\mathcal{R}0$ applies, and it will get oriented as a collider in H. Otherwise, it forms a triangle in H. If b is a non-collider on π, then it can only point to nodes in the same strongly connected component of G. But in that case, $a \ast\!\!-\!\!\ast\, b \ast\!\!-\!\!\ast\, c$ is an inducing walk between a and c in G. Hence, a, b, c will form a triangle in H. Therefore, $v \in \mathrm{PossibleDSep}_H(i, j)$. □

JM: Perhaps also show:

Lemma 9.8.5 (Tom Claassen)**.** *In a graph G, X and Y are d-separated given Z, if and only if X and Y are d-separated given Z in the graph P* obtained from the completed PAG P(G) by replacing all circle marks in P by tail marks.*

Appendix

A. Measure Theoretic Probability

1.1. Why Measure Theory?

Discrete and absolute continuous distributions are not general enough

Example 1.1.1 (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 $\Delta := \{(x,x) \,|\, x \in [0,1]\}$ 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:

$$P((X,Y) \in D) = P(\{x \in [0,1] \,|\, (x,x) \in D\}),$$

where P on the right here denotes the uniform distribution on $[0,1]$.

Notation 1.1.2 (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 1.1.3. *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 1.1.4. *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 1.1.5 (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 $A \in 2^{[0,1]}$, and:*
2. *$\lambda([a,b]) = b - a$ for all $a \leq b \in [0,1]$.*

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 1.1.6 (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. For $q \in \mathbb{Q}$ *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 1.1.7 (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* $A_1, \ldots, A_K$ *(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).*

$\Longrightarrow$ **Measure theory is the unifying 'safe space' for probability theory!**

[31] https://en.wikipedia.org/wiki/Banach-Tarski_paradox

Figure 15: Illustration of the Banach-Tarski paradox.[31]

1.2. Core Concepts

Motivation 1.2.1. *As discussed before in remark 1.1.4, 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: $P(A^c) = 1 - P(A)$ *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^c 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 1.1.5 and 1.1.7, 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 1.2.2 (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:* $A^c := \mathcal{W} \setminus A \in \mathcal{B}$,

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 1.2.3 (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.

Remark 1.2.4 (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 1.2.5 (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 1.2.6 (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 1.2.7 (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 1.2.8 (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: $f^{-1}(B) := \{w \in \mathcal{W} \mid f(w) \in B\}$.

Definition 1.2.9 (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 1.2.10 (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 1.2.11 (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 1.2.12 (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.:*

$$\{x \in \mathcal{X} \mid \neg C(x)\} \quad \text{is a } \mu\text{-null set.}$$

1.3. 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 1.3.1 (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 $\{1, \ldots, N\}$) 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 1.3.2 (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 1.3.3 (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 1.3.4. *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 1.3.5 (Borel σ-algebra on $\mathbb{R}^D$)**.** *The Borel σ-algebra of $\mathbb{R}^D$ is generated by the cubes:*

$$\mathcal{B}_{\mathbb{R}^D} = \sigma\left(\{[a_1, b_1] \times \cdots \times [a_D, b_D] \mid a_d, b_d \in \mathbb{Q}, a_d \leq b_d, d = 1, \ldots, D\}\right).$$

Definition/Lemma 1.3.6 (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 1.3.7 (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 1.3.8 (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 1.3.9 (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 1.3.10 (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).

1.4. Standard Measurable Spaces

Definition 1.4.1 (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:*

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

'Measurably isomorphic' means that there is a measurable mapping that has a measurable inverse.

Theorem 1.4.2 (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 1.4.3. *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.*

1.5. Measure Integrals

The construction of the measure integral $\int f\,d\mu$ follows in several steps.

Construction 1.5.1 (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:*

$$f_+ := \max(f, 0) \ge 0, \qquad f_- := \max(-f, 0) \ge 0.$$

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 1.5.2 (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 1.5.3 (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 1.5.4 (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 1.5.5 (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 1.5.6. *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:

$$\begin{aligned} f(x) &= f(x) \cdot \mathbb{1}_B(x) + f(x) \cdot \mathbb{1}_{B^\complement}(x) \\ &= f(w) \cdot \mathbb{1}_B(x) + f(x) \cdot \mathbb{1}_{B^\complement}(x). \end{aligned}$$

Since $w \notin B^\complement$ we get $\delta_w(B^\complement) = 0$ and thus $\int f(x) \cdot \mathbb{1}_{B^\complement}(x) \, \delta_w(dx) = 0$. Together we get:

$$\begin{aligned} \int f(x) \, \delta_w(dx) &= \int \left(f(w) \cdot \mathbb{1}_B(x) + f(x) \cdot \mathbb{1}_{B^\complement}(x)\right) \, \delta_w(dx) \\ &= f(w) \cdot \int \mathbb{1}_B(x) \, \delta_w(dx) + \underbrace{\int f(x) \cdot \mathbb{1}_{B^\complement}(x) \, \delta_w(dx)}_{=0} \\ &= f(w) \cdot \delta_w(B) \\ &= f(w). \end{aligned}$$

□

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:

$$\mathbb{E}[f] = \int f(x)\, P(dx) = \int f(x) \sum_{w \in \mathcal{W}} p(w) \cdot \delta_w(dx) = \sum_{w \in \mathcal{W}} f(w) \cdot p(w).$$

1.6. Densities/Derivatives

Definition 1.6.1. *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 1.6.2. *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 1.6.3 (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 1.6.4 (Radon-Nikodým, see [?] 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 1.6.5 (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 1.8.1) 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 $x \in \mathbb{R}^D$.

1.7. Conditional Expectation

You may be familiar with the conditional expectation for discrete random variables X, Y:

$$\mathbb{E}[X|Y=y] = \sum_{x\in\mathcal{X}} x \cdot P(X=x|Y=y) \qquad = \sum_{x\in\mathcal{X}} x \cdot \frac{P(X=x,Y=y)}{P(Y=y)}$$
$$= \frac{\sum_{x\in\mathcal{X}} x \cdot P(X=x,Y=y)}{P(Y=y)},$$

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 1.7.1 (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 1.7.2. *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 1.7.3. *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 1.7.4. *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 2.5.6): 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).*

1.8. The Lebesgue Measure

Theorem 1.8.1 (Lebesgue measure)**.** *Consider* $\mathbb{R}^D$ *with its Borel* σ*-algebra* $\mathcal{B}_{\mathbb{R}^D}$*. Then there exists a unique measure* λ^D *on* $(\mathbb{R}^D, \mathcal{B}_{\mathbb{R}^D})$ *such that:*

$$\lambda^D\left([a_1, b_1] \times \cdots \times [a_D, b_D]\right) = (b_1 - a_1) \cdots (b_D - a_D)$$

for all $a_d, b_d \in \mathbb{R}$*,* $a_d \leq b_d$*,* $d = 1, \ldots, D$*. It will be called the* D*-dimensional* Lebesgue measure*. If the dimension is clear from the context we might just write* λ *for* λ^D*.*

Theorem 1.8.2. *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 1.8.3 (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:*

$$F'(x) = f(x).$$

1.9. Transformation Rules

Theorem 1.9.1 (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 1.9.2 (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 1.9.3 (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 1.9.4 (Transformation of (probability) densities w.r.t. the Lebesgue measure)**.** *Let the setting be like in 1.9.3. 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 1.9.5 (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 $\mathcal{X}$[32]. 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.*

[32] 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.

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 1.9.6 (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:

$$\frac{d\varphi_* \mu}{d\#_{\varphi(\mathcal{W})}}(y) = \sum_{w \in \varphi^{-1}(y) \cap \mathcal{W}} m(w).$$

Example 1.9.7 (Linear transformation of Gaussian distributions)**.**

Example 1.9.8 (Density of Chi-square distributions)**.**

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. *arXiv.org preprint*, arXiv:1611.06221 [stat.ME], 2021. Accepted to The Annals of Statistics.

[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.

[Bog07] Vladimir I. Bogachev. *Measure Theory*, volume 1+2. Springer, 2007.

[CD17] Panayiota Constantinou and A. Philip Dawid. Extended Conditional Independence and Applications in Causal Inference. *The Annals of Statistics*, pages 2618–2653, 2017.

[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.

[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. *Ann. Math. Artif. Intell.*, 32(1-4):335–372, 2001.

[Daw02] A. Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70:161–189, 2002.

[Eva16] Robin J. Evans. Graphs for Margins of Bayesian Networks. *Scandinavian Journal of Statistics*, 43(3):625–648, 2016.

[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.

[Kec95] Alexander S. Kechris. *Classical Descriptive Set Theory*, volume 156 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.

[Kle14] Achim Klenke. *Probability Theory - A Comprehensive Course*. Universitext. Springer, London, 2nd edition, 2014.

[Lau96] Steffen L. Lauritzen. *Graphical Models*, volume 17. Clarendon Press, 1996.

[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.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[LZ59] Z. A. Lomnicki and S. K. Zaremba. The asymptotic distributions of estimators of the amount of transmitted information. *Information and Control*, 2:260–284, 1959.

[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.

[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.

[RERS17] Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov Properties for Acyclic Directed Mixed Graphs. *arXiv.org preprint*, arXiv:1701.06686 [stat.ME], 2017. URL: https://arxiv.org/abs/1701.06686.

[Ric03] Thomas S. Richardson. Markov Properties for Acyclic Directed Mixed Graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

[RS02] Thomas S. Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, August 2002.

[SGS00] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

[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.

[Spr20] Peter J.C. Spreij. *Measure Theoretic Probability*. UvA Course Notes, 2020. URL: https://www.science.uva.nl/~spreij/onderwijs/master/mtp.pdf.

[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.

[Wag20] Stefan Wager. Stats 361: Causal inference. Technical report, Stanford University, 2020. URL: https://web.stanford.edu/~swager/stats361.pdf.

[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.