Causal Statistical Decision Theory|What are interventions?
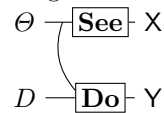
David Johnston

August 27, 2020

Contents

Note on terminology: I am trying the name "See-Do model" to describe the folowing:

I was calling it a "causal theory" before. Reasons for the change: I think "See-Do" helps to understand what the model does, and the name doesn't make premature claims to explain causality. Also, it's only two syllables which I like.

1 Notation

- X is a random variable, X is its codomain and $\mathcal{X}$ is the σ-algebra on X

- Bold letters $\mathbf{X}$ may be used for product spaces, $\mathbf{x}$ for elements of product spaces, $\mathbf{f}$ for vector valued functions and $\boldsymbol{\mathsf{X}}$ for random variables taking values in product spaces. The absence of bold font does not imply the absence of product space structure

- Nodes in a graph are italic sans serif X

- Given an indexed product space $\prod_{i \in \mathcal{I}} X_i$, $\pi_i : \prod_{i \in \mathcal{I}} X_i \to X_i$ is the projection map $(x_1, ..., x_i, ..., x_n) \mapsto x_i$

- Given a product space $X \times Y$, $\pi_X : X \times Y \to X$ is the projection map $(x, y) \mapsto x$

- Given an index set $\mathcal{I}$, $\mathbf{X}_{\mathcal{I}}$ is the indexed product space $\prod_{i \in \mathcal{I}} X_i$

- $[n]$ is the index set $\{1, ..., n\}$ for $n \in \mathbb{N}$

- Given an indexed product space $\mathbf{X}_{[n]}$, $\mathbf{X}_{<j}$ is the set $\prod_{i=1}^{j} X_i$

- $\underline{\otimes}$ is the coupled tensor product, see Definition 3.9

- Given a random variable X, F_{X} is the associated Markov kernel, see Definition 3.1

- See section 3.4.2 for rules of string diagram manipulation

- $*_X : X \to \Delta(\{*\})$ is the discard map defined in Equation 81

2 What is the difference between Causal Bayesian Networks and See-Do models?

2.1 Influence Diagrams vs See-Do models

Influence diagrams are used to represent causal models in a manner similar to, but not quite the same as, Causal Bayesian Networks. Using the version found in Dawid (2002), an influence diagram is a directed acyclic graph (DAG) with two node types: "chance" nodes and "decision" nodes. For example:

$$\boxed{D} \rightarrow \!\!\bigcirc\!\! F \tag{1}$$

Is an influence diagram with the decision node D and the chance node F. When nodes are associated with sets representing possible values, influence diagrams represent sets of Markov kernels. For example, if we associate the measurable set $(D, \mathcal{D})$ with D and $(F, \mathcal{F})$ with F, then we could take Diagram 1 to

represent the set of all Markov kernels $D \to \Delta(\mathcal{F})$, or if we add some additional assumptions it might represent a particular subset $S \subset \Delta(\mathcal{F})^D$ of these kernels that share certain properties. Compare this to a string diagram:

$$D -\boxed{\mathbb{K}}- F \tag{2}$$

This diagram represents a *particular* Markov kernel $\mathbb{K} : D \to \Delta(\mathcal{F})$. A set such as that represented by Diagram 1 could be constructed by creating a set Θ and a function $f : \Theta \to S$ that indexes each element of S with $\theta \in \Theta$. Then $S = \{f(\theta)|\theta \in \Theta\}$. If f is measurable then there is a Markov kernel $\mathbb{T} : \Theta \times D \to \Delta(\mathcal{F})$ such that $f(\theta) = \mathbb{T}_{\theta,\_}$. In this sense, for any additional assumptions that are combined with Diagram 1 to yield the set S of Markov kernels, there exists a single Markov kernel $\mathbb{T} : \Theta \times D \to \Delta(\mathcal{F})$ that generates S. $\mathbb{T}$ can be drawn as

$$\begin{array}{c} \Theta \\ D \end{array} \!\!\rangle\!\boxed{\mathbb{T}}- F \tag{3}$$

Diagram 3 has a few more elements that Diagram 1 - Θ and $\mathbb{T}$ in particular. If I were to define exactly what D, Θ, $\mathbb{T}$ and F were, Diagram 3 would represent a unique Markov kernel. On the other hand, Diagram 1 along with a precise definition of D and F could represent many different sets of Markov kernels, depending on whatever additional properties I want elements of S to share. Loosely, we can say that for any kernel $\mathbb{T}$, there is a diagram 1 along with additional assumptions that represents "the same thing".

Influence diagrams *as typically used in causal modelling* usually additional assumptions. For example, Dawid (2002) proposes that D contains a special "do nothing" element $o \in D$ such that the observations in state θ are given by $\mathbb{T}_{\theta,o}$ and consequences in state θ are given by $\mathbb{T}_{\theta,\_}$. This corresponds to the See-Do model

$$\begin{array}{c} \Theta \\ D \end{array} \!\!\!\!\begin{array}{c} \boxed{\mathbb{T}_{\_,o}}- F \\ \boxed{\mathbb{T}}- F \end{array} \tag{4}$$

While this is a feature of influence diagrams in Dawid (2002), *in general* a diagram like 1 can represent an arbitrary set of typed Markov kernels. See-Do models generate Θ-indexed sets of Markov kernels. We can therefore represent generic See-Do models with influence diagrams.

Concretely, given the influence diagram

$$\mathcal{I} = \begin{array}{c} \boxed{D} \!\rightarrow\! \boxed{F} \\ \;\; E \end{array} \tag{5}$$

and any See-Do model $\mathbb{T} : \Theta \times D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ there exists a set of auxhiliary conditions A such that the model $(\mathcal{I}, A)$ is equivalent to $\mathbb{T}$. To illustrate how influence diagrams and string diagrams compare, $\mathbb{T}$ can be drawn:

subject to irrelevance of index Θ

$$\mathbb{T} := \begin{array}{c} \Theta \;\rlap{\raise2pt{\hbox{$-\boxed{\mathbb{H}}- F$}}}\lower4pt{\hbox{$-\boxed{\mathbb{C}}- E$}} \\ D \end{array} \tag{6}$$

Maybe move the following to an appendix?

Definition 2.1 (Markov kernel/influence diagram compatibility)**.** Given a Markov kernel $\mathbb{K} : E \to \Delta(\mathcal{F})$, an influence diagram $\mathcal{I} = (S, A, E)$ and an injective $f : X \cup A \to \mathcal{F} \otimes \mathcal{E}$ which assigns each node to exactly one random variable in $\mathcal{F} \otimes \mathcal{E}$, if for all $X_1, X_2 \in \mathbf{X}$ we have $X_1 \perp_{\mathcal{I}} X_2 \implies f(X_1) \perp\!\!\!\perp_K f(X_2)$

still need to *define* conditional independence WRT kernels, though I think I can do so

Consider an arbitrary See-Do model $\mathbb{T} : D \times \Theta \to \Delta(\mathcal{E}_1 \otimes \mathcal{E}_2)$ and random variables $\mathsf{D} := \pi_D, \mathsf{E} := \pi_E, \mathsf{F} := \pi_F$ on $\mathcal{D} \otimes \mathcal{E} \otimes \mathcal{F}$. For any $\theta \in \Theta$, $\mathbb{T}_\theta$ is compatible with the influence diagram $\mathcal{I} = (\{\mathsf{E}, \mathsf{F}\}, \{\mathsf{D}\}, \{\mathsf{D} \rightarrow \mathsf{F}\})$ with respect to the injective function

$$f : \begin{cases} A \mapsto \mathsf{D} \\ E \mapsto \mathsf{E} \\ F \mapsto \mathsf{F} \end{cases} \tag{7}$$

There is always some $\mathbb{H} : \Theta \to \Delta(\mathcal{E})$ and $\mathbb{C} : \Theta \times D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ such that: $\mathbb{T}_\theta$ is equal to

$$\mathbb{T}_\theta = \begin{array}{c} \boxed{\mathbb{H}_\theta}\!-\! \mathsf{E} \\ \mathsf{D} -\!\boxed{\mathbb{C}_\theta}\!-\! \mathsf{F} \end{array} \tag{8}$$

Which implies $\mathsf{E} \perp\!\!\!\perp_{\mathbb{T}_\theta} \mathsf{F}$ and $\mathsf{E} \perp\!\!\!\perp_{\mathbb{T}_\theta} \mathsf{D}$.
The influence diagram

$$\mathcal{I} = \begin{array}{c} \boxed{D} \!\rightarrow\! \boxed{F} \\ \;\; E \end{array} \tag{9}$$

Features the d-separations $\mathsf{E} \perp_{\mathcal{I}} \mathsf{F}$ and $\mathsf{E} \perp_{\mathcal{I}} \mathsf{D}$(Peters et al., 2017; Woodward, 2016; Dawid, 2002). Thus $\mathbb{T}_\theta$ is compatible with $\mathcal{I}$ for all $\theta \in \Theta$.

2.2   Influence diagrams vs Causal Bayesian Networks

See-Do models can always be represented by influence diagrams with auxhiliary assumptions. We can then learn something about how Causal Bayesian Networks compare to See-Do models by asking how they compare to influence diagrams. The key difference between Causal Bayesian Networks and influence diagrams is that the diagrams do not contain decision nodes. Instead of Diagram 1, a Causal Bayesian Network for the same system might be

$$\boxed{F} \tag{10}$$

The difference between Diagram 1 and Diagram 10 is that the former demands a set D to be bound to the decision node D and a set F to be bound to F, while Diagram 10 demands only F. Instead of explicitly representing decisions that can be chosen, Causal Bayesian Networks *by default* feature a set of *do-interventions* on the chance nodes which seem to have a role similar to decisions in influence diagrams and See-Do models (in fact, Pearl (2009) pg 108 suggests that do-interventions and decisions are the same thing). This default set of do-interventions is what allows CBNs to avoid explicitly requiring a set D. If a set of decisions is required that is not equivalent to the set of do-interventions, this can be specified via auxhiliary assumptions, although in practice influence diagrams are usually adopted such as in (Yang et al., 2018) .

> some more examples

Dealing with a set of decisions D can be troublesome. It can easily be the case that, for example, I might be tasked with inferring a consequence map that someone else might use and I am not privy to the decisions that they might be able to make. In this case, I'd need to pick a set of decisions D which I am pretty sure covers all the possibilities.

Alternatively, the set D might be unworkably large. The set of *all* the decisions you could in principle make at this moment in as much detail as you can - this is clearly something that's far too big to write down and work with in solving an inference problem.

Speculatively, the fact that Causal Bayesian Networks default to do-interventions might help with the problems of unknown or unworkably large decision sets. If any possible decision must be resolvable to do-interventions of some type, and the effects of do-interventions are well defined, then could provide a basis for partially solving decision problems while remaining ignorant of the particular set of decisions that will ultimately be selected from.

> as in, doing the inference but not picking the best option

Beyond this, causal effects as they are informally understood *seem* to refer to univeral things, not things that depend on the set of decisions one has available. While not an especially strong reason to avoid specifying D in causal models, it is a reason nonetheless.

Do-interventions, however, cannot solve these problems. If there are no restrictions on the variables that may be included in a CBN model, then as we show do-operations and equation surgery frequently produce invalid results. To ensure that any do-interventions at all are well defined, Causal Bayesian

Networks require the specification of "intervenable variables" or "basic interventions", a requirement that is analogous to the requirement of a set of decisions D in See-Do models and influence diagrams. Specifying D may be difficult, but the do-intervention paradigm provides no solutions to this difficulty; it merely sweeps it under the rug.

2.3   What is meant by "variables"?

> Not sure where to put it, but Pearl pp 162-163 puts his models on the hook for including arbitrary variables

2.4   Necessary relationships

The relationship between a person's body mass index, their weight and their height defines what body mass index is. A fundamental claim of ours is that any causal model that defines "the causal effect of body mass index" should do so without reference to any submodel that violates this definitional relationship violation of the definition. This is an important assumption, and it rests on a judgement of what causal models ought to do. I think it is quite clear that when anyone asks for a causal effect, they expect that any operations required to define the causal effect *do not change the definitions of the variables they are employing*. While theories of causality have a role in sharpening our understanding of the term *causal effect*, the thing called a "causal effect" in an SCM should still respect some of our pre-theoretic intuitions about what causal effects are or else it should be called something else. "Causal effects" that depend on redefining variables do not respect pre-theoretic intuitions about what causal effects are:

- If I ask for the "causal effect of a person's BMI", I do not imagine that I am asking what would happen if someone's BMI were defined to be something other than their weight divided by their height

- If I ask for the "causal effect of a person's weight", I do not imagine that I am asking what would happen if someone's weight were not equal to their volume multiplied by their density

- If I ask for the "causal effect of a person's weight", I also do not imagine that I am asking what would happen if their weight were not equal to the weight of fat in their body plus the weight of all non-fat parts of their body

- If I ask for the "causal effect of taking a medicine", I do not imagine that I am asking what would happen if a person were declared to have taken a medicine independently of whatever substances have actually entered their body and how they entered

We will call relationships that have to hold *necessary relationships*. We provide the example of relationships that have to hold by definition as examples,

but definitions may not be the only variety of necessary relationships. For example, one might also wish to stipulate that certain laws of physics are required to hold in all submodels.

If a causal model contains variables that are necessarily related, then an intervention on one of them must always change another variable in the relationship. If I change a person's weight, their height or BMI must change (or both). If I change their height, their weight or BMI must change and if I change their BMI then their weight or height must change. This conflicts with the usual acyclic definition of causal models, where the proposition that A causes B rules out the possibility that B or any of its descendents are a cause of A. Thus in an acyclic model it isn't possible for for an intervention on BMI to change weight or height and interventions on weight and height to also change BMI. Theroem 2.11 formalises this conflict for recursive structural causal models: for any set of variables that are necessarily related by a cyclic relationship, at least one of them has no hard interventions defined.

2.5 Recursive Structural Causal Models

We begin by showing that necessary relationships are incompatible with structural causal models.

Definition 2.2 (Recursive Structural Causal Model)**.** A recursive structural causal model (SCM) is a tuple

$$\mathcal{M} := \langle N, M, \mathbf{X}_{[N]}, \mathbf{E}_{[M]}, \{f_i | i \in [N]\}, \mathbb{P}_{\mathcal{E}} \rangle \tag{11}$$

where

- $N \in \mathbb{N}$ is the number of *endogenous variables* in the model

- $M \in \mathbb{N}$ is the number of *exogenous variables* in the model

- $\mathbf{X}_{[N]} := \{X_i | i \in [N]\}$ where, for each $i \in [N]$, $(X_i, \mathcal{X}_i)$ is a standard measurable space taking and the codomain of the i-th endogenous variable

- $\mathbf{E}_{[M]} := \{E_j | j \in [M]\}$ where, for $j \in [M]$, E_j is a standard measurable space and the codomain of the j-th exogenous variable

- $f_i : \mathbf{X}_{<i} \times \mathbf{E}_{\mathcal{J}} \to X_i$ is a measurable function which we call *the causal mechanism controlling the i-th endogenous variable*

- $\mathbb{P}_{\mathcal{E}} \in \Delta(\mathbf{E}_{\mathcal{J}})$ is a probability measure on the space of exogenous variables

Definition 2.3 (Observable kernel)**.** Given an SCM $\mathcal{M}$ with causal mechanisms $\{f_i | i \in [N]\}$, define the *observable kernel* $G_i : E \to \Delta(\mathbf{X}_{[i]})$ recursively:

$$G_1 = \quad \mathbf{E}_{[M]} - \boxed{F_{f_1}} - X_1 \quad f_1 \tag{12}$$

$$G_{n+1} = \quad \mathbf{E}_{[M]} - \boxed{G_n} \overset{\displaystyle \mathbf{X}_{<n+1}}{\underset{\boxed{F_{f_{n+1}}} - X_{n+1}}{}} \tag{13}$$

7

Definition 2.4 (Joint distribution on endogenous variables)**.** The *joint distribution on endogenous variables* defined by $\mathcal{M}$ is $\mathbb{P}_{\mathcal{M}} := \mathbb{P}_{\mathcal{E}} G_N$ (which is the regular kernel product, see Definition 3.3). For each $i \in [N]$ define the random variable $\mathsf{X}_i : \mathbf{X}_{[N]} \to X_i$ as the projection map $\pi_i : (x_1, ..., x_i, .., x_N) \mapsto x_i$. By Lemma 2.5, $\underline{\otimes}_{i \in [N]} \mathsf{X}_i = \mathrm{Id}_{\mathbf{X}_{[N]}}$, and so $\mathbb{P}_{\mathcal{M}}$ is the joint distribution of the variables $\{\mathsf{X}_i | i \in [N]\}$.

I use the notation $\mathbb{P}_{\mathcal{M}}$ rather than $\mathbb{P}_{\mathsf{X}_{[N]}}$ to emphasize the dependence on the model $\mathcal{M}$.

Lemma 2.5 (Coupled product of all random variables is the identity)**.** $\underline{\otimes}_{i \in [N]} \mathsf{X}_i = \mathrm{Id}_{\mathbf{X}_{[N]}}$

Proof. for any $\mathbf{X} \in \mathbf{X}_{[N]}$,

$$\underline{\otimes}_{i \in [N]} \mathsf{X}_i(\mathbf{X}) = (\pi_1(\mathbf{X}), ..., \pi_N(\mathbf{X})) \tag{14}$$

$$= (x_1, ..., x_n) \tag{15}$$

$$= \mathbf{X} \tag{16}$$

$\square$

Definition 2.6 (Hard Interventions)**.** Let $\mathcal{M}$ be the set of all *SCMs* sharing the indices, spaces and measure $\langle N, M, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{[M]}, \mathbb{P}_{\mathcal{E}} \rangle$. Note that the causal mechanisms are not fixed.

Given an SCM $\mathcal{M} = \langle N, M, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{[M]}, \{f_i | i \in [N]\}, \mathbb{P}_{\mathcal{E}} \rangle$ and $\mathcal{S} \subset [N]$, a *hard intervention* on $\mathsf{X}_{\mathcal{S}}$ is a map $Do_{\mathcal{S}} : \mathbf{X}_{\mathcal{S}} \times \mathcal{M} \to \mathcal{M}$ such that for $\mathbf{a} \in \mathbf{X}_{\mathcal{S}}$, $Do_{\mathcal{S}}(\mathbf{a}, \mathcal{M}) = \langle N, M, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{[M]}, \{f_i' | i \in [N]\}, \mathbb{P}_{\mathcal{E}} \rangle$ where

$$f_i' = f_i \qquad\qquad i \notin \mathcal{S} \tag{17}$$

$$f_i' = \pi_i(\mathbf{a}) \qquad\qquad i \in \mathcal{S} \tag{18}$$

To match standard notation, we will write $\mathcal{M}^{do(\mathsf{X}_{\mathcal{S}} = \mathbf{a})} := Do_{\mathcal{S}}(\mathbf{a}, \mathcal{M})$

2.6 Recursive Structural Causal Models with Necessary Relationships

Necessary relationships are extra constraints on the joint distribution on endogenous variables defined by an SCM. For example, given an SCM $\mathcal{M}$ if the variable X_1 represents weight, X_2 represents height and X_3 represents BMI then we want to impose the constraint that

$$\mathsf{X}_3 = \frac{\mathsf{X}_1}{\mathsf{X}_2} \tag{19}$$

$\mathbb{P}_{\mathcal{M}}$-almost surely.

Definition 2.7 (Constrained Recursive Structural Causal Model (CSCM))**.** A CSCM $\mathcal{M} := \langle N, M, \mathbf{X}_{[N]}, \mathbf{E}_{[M]}, \{f_i | i \in [N]\}, \{r_i | i \in [N]\}, \mathbb{P}_{\mathcal{E}} \rangle$ is an SCM along with a set of *constraints* $r_i : \mathbf{X}_{[N]} \to X_i$.

If $\mathsf{X}_i = r_i(\mathsf{X}_{[N]})$ $\mathbb{P}_{\mathcal{M}}$-almost surely then M is *valid*, otherwise it is *invalid*.

8

We can recover regular SCMs by imposing only trivial constraints:

Lemma 2.8 (CSCM with trivial constraints is always valid). *Let $\mathcal{M}$ be a CSCM with the trivial constraints $r_i = \pi_i$ for all $i \in [N]$. Then $\mathcal{M}$ is valid.*

Proof. By definition 2.7, we require $\mathsf{X}_i = \mathsf{X}_i$, $\mathbb{P}_{\mathcal{M}}$-almost surely. $\mathsf{X}_i(\mathbf{X}) = \mathsf{X}_i(\mathbf{X})$ for all $\mathbf{X} \in \mathbf{X}_{[N]}$ and $P_{\mathcal{M}}(\mathbf{X}_{[N]}) = 1$, therefore $\mathcal{M}$ is valid. $\qquad\square$

Call a constraint r_i *cyclic* if $\mathsf{X}_i = r_i(\mathsf{X}_{[N]})$ implies there exists an index set $O \subset [N]$, $O \ni i$, such that for each $j \in O$, $\mathbf{b} \in \mathbf{X}_{O \setminus \{j\}}$ there exists $a \in X_j$ such that

$$\mathsf{X}_{O \setminus \{j\}} = \mathbf{b} \tag{20}$$
$$\implies \mathsf{X}_j = a \tag{21}$$

BMI is an example of a cyclic constraint if we insist that weight and height are always greater than 0. If $\mathsf{X}_3 = \frac{\mathsf{X}_1}{\mathsf{X}_2}$ then we have:

$$[\mathsf{X}_1, \mathsf{X}_2] = [b_1, b_2] \tag{22}$$
$$\implies \mathsf{X}_3 = \frac{b_1}{b_2} \tag{23}$$
$$[\mathsf{X}_2, \mathsf{X}_3] = [b_2, b_3] \tag{24}$$
$$\implies \mathsf{X}_1 = b_2 b_3 \tag{25}$$
$$[\mathsf{X}_1, \mathsf{X}_3] = [b_1, b_3] \tag{26}$$
$$\implies \mathsf{X}_2 = \frac{b_1}{b_3} \tag{27}$$

<div style="background-color:orange">
The following is a generally useful lemma that should probably be in basic definitions of Markov kernel spaces
</div>

Lemma 2.9 (Projection and selectors). *Given an indexed product space $\mathbf{X} := \prod_{i \in \mathcal{I}} X_i$ with ordered finite index set $\mathcal{I} \ni i$, let $\pi_i : \mathbf{X} \to X_i$ be the projection of the i-indexed element of $\mathbf{X} \in \mathbf{X}$.*

Let $F_{\pi_i} : \mathbf{X} \to \Delta(\mathcal{X}_i)$ be the Markov kernel associated with the function π_i, $F_{\pi_i} : \mathbf{X} \mapsto \delta_{\pi_i(\mathbf{X})}$. Given $O \subset \mathcal{I}$, define the selector S_i^O:

$$S_i^O = \begin{cases} \mathrm{Id}_{X_i} & i \in O \\ \maltese_{X_i} & i \notin O \end{cases} \tag{28}$$

Then $\underline{\otimes}_{i \in O} F_{\pi_i} = \otimes_{i \in \mathcal{I}} S_i^O$.

Proof. Suppose O is the empty set. Then the empty tensor product $\otimes_{i \in \emptyset} S_i$ and the empty coupled tensor product $\underline{\otimes}_{i \in \emptyset} F_{\pi_i}$ are both equal to $\maltese_{\mathbf{X}}$.

By definition of F_{π_i}, $F_{\pi_i} = \otimes_{i \in \mathcal{I}} S_i^{\{i\}}$.

Suppose for $P \subsetneq O$ with greatest element k we have $\underline{\otimes}_{i \in P} F_{\pi_i} = \otimes_{i \in \mathcal{I}} S_i^P$, and suppose that j is the next element of O not in P.

$$(\underline{\otimes}_{i \in P} F_{\pi_i}) \underline{\otimes} F_{\pi_j} = \quad \mathbf{X} \Big\langle \begin{array}{l} \boxed{\underline{\otimes}_{i \in P} F_{\pi_i}} - \mathbf{X}_P \\ \boxed{F_{\pi_j}} - X_j \end{array} \tag{29}$$

$$= \quad \begin{array}{l} \mathbf{X}_{<j} \\ X_j \\ \mathbf{X}_{>j} \end{array} \begin{array}{l} \boxed{\underline{\otimes}_{i \in P} F_{\pi_i}} - \mathbf{X}_P \\ \boxed{F_{\pi_j}} - X_j \end{array} \tag{30}$$

$$= \quad \begin{array}{l} \mathbf{X}_{<j} \\ X_j \\ \mathbf{X}_{>j} \end{array} \begin{array}{l} \boxed{\underline{\otimes}_{i \in P} F_{\pi_i}} - \mathbf{X}_P \\ \ast \\ \ast \quad - X_j \end{array} \tag{31}$$

$$= \quad \begin{array}{l} \mathbf{X}_{<j} \\ X_j \\ \mathbf{X}_{>j} \end{array} \begin{array}{l} \boxed{\underline{\otimes}_{i \in P} F_{\pi_i}} - \mathbf{X}_P \\ X_j \end{array} \tag{32}$$

$$= \quad \begin{array}{l} \mathbf{X}_{<j} \\ X_j \\ \mathbf{X}_{>j} \end{array} \begin{array}{l} \boxed{\otimes_{i \in \mathcal{I}} S_i^P} - \mathbf{X}_P \\ X_j \end{array} \tag{33}$$

Because all elements of P are less than j, the selector S_k^P resolves to the discard map for $k > j$:

$$= \quad \begin{array}{l} \mathbf{X}_{<j} \\ X_j \\ \mathbf{X}_{>j} \end{array} \begin{array}{l} \ast \boxed{\otimes_{i<j} S_i^P} - \mathbf{X}_P \\ X_j \end{array} \tag{34}$$

$$= \quad \begin{array}{l} \mathbf{X}_{<j} \\ X_j \\ \mathbf{X}_{>j} \end{array} \begin{array}{l} \boxed{\otimes_{i<j} S_i^P} - \mathbf{X}_P \\ X_j \\ \ast \end{array} \tag{35}$$

$$= \quad \begin{array}{l} \mathbf{X}_{<j} \\ X_j \\ \mathbf{X}_{>j} \end{array} \boxed{\otimes_{i \in \mathcal{I}} S_i^{P \cup \{j\}}} \begin{array}{l} \mathbf{X}_P \\ X_j \end{array} \tag{36}$$

Where 36 follows from the definition of the selector $S_i^{P \cup \{j\}}$.
The proof follows by induction on the elements of O.

$\square$

Lemma 2.10 (Hard interventions do not affect the joint distributions of earlier variables). *Given a CSCM $\mathcal{M} = \langle N, M, \mathbf{X}_{[N]}, \mathbf{E}_{[M]}, \{f_i | i \in [N]\}, \{r_i | i \in [N]\}, \mathbb{P}_{\mathcal{E}}\rangle$, any $k \in [N]$ and any $O \subset [k-1]$, $P_{\mathcal{M}}(\mathsf{X}_O) = P_{\mathcal{M}}^{do(\mathsf{X}_k)=a}(\mathsf{X}_O)$ for all $a \in X_k$.*

Proof. Let $G_i^{\mathcal{Q}}$, $i \in [N]$ be the i-th iteration of the kernel defined in Equations 12 and 13 with respect to model $\mathcal{Q}$. Note that from Equation 13

$$\mathbf{E}_{[M]} - \boxed{G_i^{\mathcal{Q}}} - \mathbf{X}_{<i} \;\; = \;\; \mathbf{E}_{[M]} - \boxed{G_{i-1}^{\mathcal{Q}}} - \boxed{F_{f_i}} - \mathbf{X}_{<i} \tag{37}$$

$$= G_{i-1}^{\mathcal{Q}} \tag{38}$$

It follows that

$$\mathbf{E}_{[M]} - \boxed{G_N} - \mathbf{X}_{<i} \;\; = G_{i-1} \tag{39}$$

Because $f_i = f_i^{do(\mathsf{X}_k=a)}$ for $i < k$, we have

$$G_i^{\mathcal{M}} = G_i^{\mathcal{M}^{do(\mathsf{X}_k=a)}} \tag{40}$$

for all $i < k$. By lemma 2.9, for any $O \subset [k-1]$ we have $F_{\mathsf{X}_O} = \otimes_{i\in[N]}S_i^O$. As there are no elements of O greater than or equal to k, the selector S_i^O resolves to the discard for all $i >= k$. Thus $F_{\mathsf{X}_O} = (\otimes_{i\in[k-1]}S_i^O) \otimes \ast_{\mathbf{X}_{[N]\setminus[k-1]}}$. Defining $S_{[k-1]}^O := \otimes_{i\in[k-1]}S_i^O$, we have:

$$F_{\mathsf{X}_O} = \;\; \mathbf{X}_{[k-1]} - \boxed{S_{[k-1]}^O} - \mathbf{X}_O \qquad \mathbf{X}_{[N]\setminus[k-1]} - \ast \tag{41}$$

Thus

$$\mathbb{P}_{\mathcal{M}}(\mathsf{X}_O) = \mathbb{P}_{\mathcal{E}} G_N^{\mathcal{M}} F_{\mathsf{X}_O} \tag{42}$$

$$\overset{41}{=} \; \langle\mathbb{P}_{\mathcal{E}}| \; \boxed{G_N^{\mathcal{M}}} \; \boxed{S_{[k-1]}^O} - \mathbf{X}_O \quad \ast \, (\mathbf{X}_{[N]\setminus[k-1]}) \tag{43}$$

$$\overset{39}{=} \; \langle\mathbb{P}_{\mathcal{E}}| \; \boxed{G_{k-1}^{\mathcal{M}}} \; \boxed{S_{[k-1]}^O} - \mathbf{X}_O \tag{44}$$

$$\overset{40}{=} \; \langle\mathbb{P}_{\mathcal{E}}| \; \boxed{G_{k-1}^{\mathcal{M}^{do(\mathsf{X}_k=a)}}} \; \boxed{S_{[k-1]}^O} - \mathbf{X}_O \tag{45}$$

$$\overset{39}{=} \; \langle\mathbb{P}_{\mathcal{E}}| \; \boxed{G_N^{\mathcal{M}^{do(\mathsf{X}_k=a)}}} \; \boxed{S_{[k-1]}^O} - \mathbf{X}_O \quad \ast \, (\mathbf{X}_{[N]\setminus[k-1]}) \tag{46}$$

$$= P_{\mathcal{M}^{do(\mathsf{X}_k=a)}}(\mathsf{X}_O) \tag{47}$$

11

$\square$

Theorem 2.11 (Undefined hard interventions with cyclic constraints)**.** *Consier a CSCM $\mathcal{M} = \langle N, M, \mathbf{X}_{[N]}, \mathbf{E}_{[M]}, \{f_i | i \in [N]\}, \{r_i | i \in [N]\}, \mathbb{P}_{\mathcal{E}} \rangle$ with r_i a cyclic constraint with respect to $O \subset [N]$ and the rest of the constraints trivial: $r_j = \pi_j$, $j \neq i$, and suppose $\mathcal{M}$ is valid.*

If for each $k \in O$, $\exists A \in \mathcal{X}_i$ such that $0 < \mathbb{P}_{\mathcal{M}}(\mathsf{X}_i \in A) < 1$ then for at least one $k \in O$ all models given by a hard intervention on X_k are invalid.

Proof. Choose k to be the maximum element of O. By the assumption $\mathcal{M}$ is valid, we have $\mathsf{X}_i = r_i(\mathbf{X})$, $\mathbb{P}_{\mathcal{M}}$-almost surely. Let $B^A = \{\mathbf{b} \in \mathbf{X}_{O \backslash k} | \mathsf{X}_{O \backslash k} = \mathbf{b} \implies \mathsf{X}_k \in A\}$ and $B^{A^C} = \{\mathbf{b} \in \mathbf{X}_{O \backslash k} | \mathsf{X}_{O \backslash k} = \mathbf{b} \implies \mathsf{X}_k \notin A\}$.

r_i holds on a set of measure 1, and wherever it holds $\mathsf{X}_{O \backslash \{k\}}$ is either in B^A or B^{A^C}. Thus $\mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O \backslash \{k\}} \in B^A \cup B^{A^C}) = 1$.

B^A and B^{A^C} are disjoint.

By construction, $\mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O \backslash \{k\}} \in B^A \ \& \ \mathsf{X}_k \in A) = \mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O \backslash \{k\}} \in B^A)$ and $\mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O \backslash \{k\}} \in B^{A^C} \ \& \ \mathsf{X}_k \in A^C) = \mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O \backslash \{k\}} \in B^{A^C})$

By additivity, $\mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O \backslash \{k\}} \in B^A \ \& \ \mathsf{X}_k \in A) + \mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O \backslash \{k\}} \notin B^A \ \& \ \mathsf{X}_k \in A) = P_{\mathcal{M}}(\mathsf{X}_k \in A)$.

By additivity agian

$$\mathbb{P}_{\mathcal{M}}\left(\mathsf{X}_{O \backslash \{k\}} \notin B^A \ \& \ \mathsf{X}_k \in A\right) = \mathbb{P}_{\mathcal{M}}\left(\mathsf{X}_{O \backslash \{k\}} \in B^{A^C} \ \& \ \mathsf{X}_k \in A\right) \tag{48}$$
$$+ \mathbb{P}_{\mathcal{M}}\left(\mathsf{X}_{O \backslash \{k\}} \in (B^{A^C} \cup B^A)^C \ \& \ \mathsf{X}_k \in A\right) \tag{49}$$
$$<= 0 + P_{\mathcal{M}}\left(\mathsf{X}_{O \backslash \{k\}} \in (B^{A^C} \cup B^A)^C\right) \tag{50}$$
$$= 0 \tag{51}$$

Thus $\mathbb{P}_{\mathcal{M}}\left(\mathsf{X}_{O \backslash \{k\}} \in B^A \ \& \ \mathsf{X}_k \in A\right) = P_{\mathcal{M}}(\mathsf{X}_k \in A) = P_{\mathcal{M}}(\mathsf{X}_{O \backslash \{k\}} \in B^A)$ and by an analogous argument $\mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O \backslash \{k\}} \in B^{A^C}) = P_{\mathcal{M}}(\mathsf{X}_k \in A^C)$.

Choose some $a \in A$, and consider the hard intervention $\mathcal{M}^{do(\mathsf{X}_k = a)}$. Suppose $\mathcal{M}^{do(\mathsf{X}_k = a)}$ is also valid. Then, as before, $\mathbb{P}_{\mathcal{M}^{do(\mathsf{X}_k = a)}}(\mathsf{X}_{O \backslash \{k\}} \in B^{A^C}) = \mathbb{P}_{\mathcal{M}^{do(\mathsf{X}_k = a)}}(\mathsf{X}_k \in A^C)$.

By definition of hard interventions, $f_k^{do(\mathsf{X}_k = a)} = a$. Thus $G_N^{\mathcal{M}^{do(\mathsf{X}_k = a)}} F_{\mathsf{X}_k}$ is the kernel $\mathbf{X} \mapsto \delta_a$ and it follows that $\mathbb{P}_{\mathcal{M}^{do(\mathsf{X}_k = a)}}(\mathsf{X}_k) = \delta_a$.

By lemma 2.10, $\mathbb{P}_{\mathcal{M}^{do(\mathsf{X}_k = a)}}(\mathsf{X}_{O \backslash \{k\}} \in B^{A^C}) = P_{\mathcal{M}}(\mathsf{X}_{O \backslash \{k\}} \in B^{A^C}) = P_{\mathcal{M}}(\mathsf{X}_k \in A^C) > 0$. But $P_{\mathcal{M}^{do(\mathsf{X}_k = a)}}(\mathsf{X}_k \in A^C) = \delta_z(\mathsf{X}_k \in A^C) = 0$, contradicting the assumption of validity of $\mathcal{M}^{do(\mathsf{X} = a)}$.

An analogous argument shows that all hard interventions $a' \in A^C$ are also invalid. $\square$

2.7 Cyclic Structural Causal Models

It is not very surprising that acyclic causal models cannot accommodate cyclic constraints. Can cyclic causal models do so? While Bongers et al. (2016) has develope a theory of cyclic causal models, cyclic are generally far less well understood than acyclic models. I show that the theory of cyclic models that Bongers has developed also fails to define hard interventions on variables subject to cyclic constraints. This does not rule out the possibility that there is some other way to define cyclic causal models that do handle these constraints, but I have not taken it upon myself to develop such a theory.

> Haven't done any work from here on

We adopt the framework of cyclic structural causal models to make our arguments, adapted from Bongers et al. (2016). This is somewhat non-standard, but allows us to make a stronger argument for the impossibility of modelling arbitrary sets of variables using structural interventional models.

Definition 2.12 (Structural Causal Model)**.** A structural causal model (SCM) is a tuple

$$\mathcal{M} := \langle \mathcal{I}, \mathcal{J}, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{\mathcal{J}}, \mathbf{f}_{\mathcal{I}}, \mathbb{P}_{\mathcal{E}}, \mathbf{\textit{E}}_{\mathcal{J}} \rangle \tag{52}$$

where

- $\mathcal{I}$ is a finite index set of *endogenous variables*

- $\mathcal{J}$ is a finite index set of *exogenous variables*

- $\mathbf{X}_{\mathcal{I}} := \{X_i\}_{\mathcal{I}}$ where, for each $i \in \mathcal{I}$, $(X_i, \mathcal{X}_i)$ is a standard measurable space taking and the codomain of the i-th endogenous variable

- $\mathbf{E}_{\mathcal{J}} := \{E_j\}_{\mathcal{J}}$ where, for $j \in \mathcal{J}$, E_j is a standard measurable space and the codomain of the j-th endogenous variable

- $\mathbf{f}_{\mathcal{I}} = \otimes_{i \in \mathcal{I}} f_i$ is a measurable function, and $f_i : \mathbf{X}_{\mathcal{I}} \times \mathbf{E}_{\mathcal{J}} \to X_i$ is the causal mechanism controlling X_i

- $\mathbb{P}_{\mathcal{E}} \in \Delta(\mathbf{E}_{\mathcal{J}})$ is a probability measure on the space of exogenous variables

- $\mathbf{\textit{E}}_{\mathcal{J}} = \otimes_{j \in \mathcal{J}} \mathsf{E}_j$ is the set of exogenous variables, with $\mathbb{P}_{\mathcal{E}} = \mathbf{\textit{E}}_{\mathcal{J}\#} P_{\mathcal{E}}$ and E_j is the j-th exogenous variable with marginal distribution given by $\mathsf{E}_{j\#} \mathbb{P}_{\mathcal{E}}$

If for $\mathbb{P}_{\mathcal{E}}$-almost every $\mathbf{e} \in \mathbf{E}_{\mathcal{J}}$ there exists $\mathbf{X} \in \mathbf{X}_{\mathcal{I}}$ such that

$$\mathbf{X} = \mathbf{f}_{\mathcal{I}}(\mathbf{X}, \mathbf{e}) \tag{53}$$

Then an SCM $\mathcal{M}$ induces a unique probability space $(\mathbf{X}_{\mathcal{I}} \times \mathbf{E}_{\mathcal{J}}, \mathcal{X}_{\mathcal{I}} \otimes \mathcal{E}_{\mathcal{J}}, \mathbb{P}_{\mathcal{M}})$ (Bongers et al., 2016). If no such solution exists then we will say an SCM is invalid, as it imposes mutually incompatible constraints on the endogenous variables. It may be also the case that multiple solutions exist.

If an SCM induces a unique probability space then there exist random variables $\{\mathsf{X}_i\}_{i \in \mathcal{I}}$ such that, $P_{\mathcal{M}}$ almost surely Bongers et al. (2016):

$$\mathsf{X}_i = f_i(\boldsymbol{X}_{\mathcal{I}}, \boldsymbol{E}_{\mathcal{J}}) \tag{54}$$

Where $\boldsymbol{X}_{\mathcal{I}} = \otimes_{i \in \mathcal{I}} \mathsf{X}_i$.

A structural causal model can be transformed by *mechanism surgery*. Given $\mathcal{S} \subset \mathcal{I}$ and a set of new functions $\mathbf{f}_{\mathcal{S}}^I : \mathbf{X}_{\mathcal{S}} \times \mathbf{E}_{\mathcal{J}} \to \mathbf{X}_{\mathcal{S}}$, mechanism surgery "replaces" the corresponding parts of $\mathbf{f}_{\mathcal{I}}$ with $\mathbf{f}_{\mathcal{S}}^I$.

Definition 2.13 (Mechanism surgery)**.** Let $\mathcal{M}$ be the set of SCMs with elements $\langle \mathcal{I}, \mathcal{J}, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{\mathcal{J}}, \_\_, \mathbb{P}_{\mathcal{E}}, \boldsymbol{E}_{\mathcal{J}} \rangle$ (note that the causal mechanisms are unspecified). Mechanism surgery is an operation $I : \mathbf{X}_{\mathcal{I}}^{\mathbf{X}_{\mathcal{I}} \times \mathbf{E}_{\mathcal{J}}} \times \mathcal{M} \to \mathcal{M}$ that takes a causal model $\mathcal{M}$ with arbitrary causal mechanisms and a set of causal mechanisms $\mathbf{f}_{\mathcal{I}}'$ and maps it to a model $\mathcal{M}' = \langle \mathcal{I}, \mathcal{J}, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{\mathcal{J}}, \mathbf{f}_{\mathcal{I}}', \mathbb{P}_{\mathcal{E}}, \boldsymbol{E}_{\mathcal{J}} \rangle$.

If $\mathcal{M}$ has causal mechanisms $\mathbf{f}_{\mathcal{I}}$ and $\mathcal{O} \subset \mathcal{I}$ is the largest set such that $\pi_{\mathcal{O}} \circ \mathbf{f}_{\mathcal{I}} = \pi_{\mathcal{O}} \circ \mathbf{f}_{\mathcal{I}}'$ then we say I is an *intervention* on $\mathcal{L} := \mathcal{I} \setminus \mathcal{O}$. We will use the special notation $\mathcal{M}^{I(\mathcal{L}), \mathbf{f}_{\mathcal{L}}'} := I(\mathcal{M}, \mathbf{f}_{\mathcal{L}}'$ to denote an SCM related to $\mathcal{M}$ by intervention on a subset of $\mathcal{I}$.

If *furthermore* $\pi_{\mathcal{L}} \mathbf{f}_{\mathcal{I}}'$ is a constant function equal to $\mathbf{a}$, then we say I is a *hard intervention* on $\mathcal{L}$. We use the special notation $\mathcal{M}^{Do(\mathcal{L})=\mathbf{a}} := I(\mathcal{M}, \mathbf{f}_{\mathcal{L}}'$ to denote SCMs related to $\mathcal{M}$ by hard interventions. We also say that the *causal effect* of $\mathcal{L}$ is the set of SCNMs $\{\mathcal{M}^{Do(\mathcal{L})=\mathbf{a}} | a \in \mathbf{X}_{\mathcal{L}}\}$.

We say a *causal model* is any kind of model that defines causal effects. An SCM $\mathcal{M}$ in combination with hard interventions defines causal effects, so an SCM is a causal model. Call each interventional model $\mathcal{M}^{do(\mathsf{X}_i = x)}$ a *submodel* of $\mathcal{M}$.

Strictly, the random variables X_i depend on the probability space induced by a particular model $\mathcal{M}$, they are intended to refer to "the same variable" across different models that are related by mechanism surgery. We will abuse notation and use X_i to refer to the *family* of random variables induced by a set of models related by mechanism surgery, and rely on explicitly noting the measure $\mathbb{P}_{...}(...)$ to specify exactly which random variables we are talking about.

In practice, we typically specify a "small" SCM containing a few endogenous variables $\mathcal{I}$ (called a "marginal SCM" by Bongers et al. (2016)) which is understood to summarise the relevant characteristics of a "large" SCM containing many variables $\mathcal{I}^*$. We will argue that without restrictions on the large set of variables $\mathcal{I}^*$, surgically transformed SCMs will usually be invalid.

> Incidentally, this messiness with random variables can be solved if we use See-Do models.

2.8 Not all variables have well-defined interventions

A long-running controversy about causal inference concerns the question of "the causal effect of body mass index on mortality". On the one hand, Hernán and Taubman (2008) and others claim that there is no well-defined causal effect of a person's body mass index (BMI), defined as their weight divided by their height,

and their risk of death. Pearl claims, in defence of Causal Bayesian Networks, that the causal effect of *obesity* is well-defined, though it is not clear whether he defends the proposition that BMI itself has a causal effect:

> That BMI is merely a coarse proxy of obesity is well taken; obesity should ideally be described by a vector of many factors, some are easy to measure and others are not. But accessibility to measurement has no bearing on whether the effect of that vector of factors on morbidity is "well defined" or whether the condition of consistency is violated when we fail to specify the interventions used to regulate those factors. (Pearl, 2018)

We argue that BMI does *not* have a well-defined causal effect, and without further assumptions neither does any variable.

2.8.1 Necessary relationships in cyclic SCMs

If an SCM contains variables that are necessarily related, we wish to impose the additional restriction that these necessary relationships hold for every submodel. This can be done by extending the previous definition:

Definition 2.14 (SCM with necessary relationships)**.** An SCM with necessary relationships (SCNM) is a tuple $\mathcal{M} := \langle \mathcal{I}, \mathcal{J}, \mathbf{X}_\mathcal{I}, \mathbf{E}_\mathcal{J}, \mathbf{f}_\mathcal{I}, \mathbf{g}_\mathcal{I}, \mathbb{P}_\mathcal{E}, \boldsymbol{E}_\mathcal{J} \rangle$, which is an SCM with the addition of a vector function of *necessary relationships* $\mathbf{g}_\mathcal{I} := \underline{\otimes}_{i \in \mathcal{I}} g_i$ where each $g_i : \mathbf{X}_\mathcal{I} \to X_i$ is a necessary relationship involving X_i.

An SCM with necessary induces a unique probability space if for $\mathbb{P}_\mathcal{E}$-almost every $e \in \mathcal{E}$ there exists a unique $\mathbf{X} \in \mathbf{X}_\mathcal{I}$ such that simultaneously

$$\mathbf{X} = \mathbf{f}_\mathcal{I}(\mathbf{X}, \mathbf{e}) \tag{55}$$

$$\mathbf{X} = \mathbf{g}_\mathcal{I}(\mathbf{X}) \tag{56}$$

If no such $\mathbf{X}$ exists then an SCNM is invalid.

Mechanism surgery for SCNMs involves modification of $\mathbf{f}_\mathcal{I}$ only, just like SCMs.

If we wish to stipulate that a particular variable X_i has no causal relationships or necessary relationships we can specify this with the trivial mechanisms $f_i : (\mathbf{X}, \mathbf{e}) \mapsto x_i$ and $g_i : \mathbf{X} \mapsto x_i$ respectively. An SCNM $\mathcal{M}$ with the trivial necessary relationship $\mathbf{g}_\mathcal{I} : \mathbf{X} \mapsto \mathbf{X}$ induces the equivalent probability spaces as the SCM obtained by removing $\mathbf{g}_\mathcal{I}$ from $\mathcal{M}$.

Because BMI is always equal height/weight, given some SCNM $\mathcal{M}$ containing endogenous variables X_h, X_w and X_b representing height, weight and BMI respectively, it should be possible to construct a more "primitive" SCNM $\mathcal{M}^p$ containing every variable $\mathcal{M}$ does except X_b that agrees with $\mathcal{M}$ on all interventions except those on X_b.

Definition 2.15 (Marginal model)**.** Given an SCNM

$$\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathbf{X}_\mathcal{I}, \mathbf{E}_\mathcal{J}, \mathbf{f}_\mathcal{I}, \mathbf{g}_\mathcal{I}, \mathbb{P}_\mathcal{E}, \boldsymbol{E}_\mathcal{J} \rangle$$

a marginal model over $\mathcal{L} \subset \mathcal{I}$ is an SCNM

$$\mathcal{M}^{*_L} = \langle \mathcal{O}, \mathcal{J}, \mathbf{X}_{\mathcal{O}}, \mathbf{E}_{\mathcal{J}}, \mathbf{f}_{\mathcal{O}}^{\mathcal{L}*}, \mathbf{g}_{\mathcal{O}}^{\mathcal{L}*}, \mathbb{P}_{\mathcal{E}}, \boldsymbol{E}_{\mathcal{J}} \rangle$$

such that $(\mathbb{P}_{\mathcal{M}})^*_{\mathcal{L}} = \mathbb{P}_{(\mathcal{M}^{*_L})}$ and for all interventions $\mathbf{f}'_{\mathcal{O}}$ on $\mathcal{O} := \mathcal{I} \setminus \mathcal{L}$ that do not depend on $\mathcal{L}$

$$(\mathbb{P}_{\mathcal{M}^{I(\mathcal{O})}, \mathbf{f}'_{\mathcal{O}}})^*_{\mathcal{L}} = \mathbb{P}_{(\mathcal{M}^{*_{\mathcal{L}}}, I(\mathcal{O}), \mathbf{f}'_{\mathcal{O}} \circ \pi_{\mathcal{O}})}$$

A *primitive model* is a special case of a marginal model where any intervention that depended only on endogenous variables in the original model can be replicated with some intervention that depends only on endogenous variables in the marginal model. If the endogenous variables represent *observed* variables, then the plausible intervention operations may only be allowed to depend on these variables. In general, there may be interventions that are possible in the original model that are not possible in the marginal model.

Definition 2.16 (Primitive model). A *primitive model* $\mathcal{M}^p$ is a marginal model of $\mathcal{M}$ with respect to some $\mathcal{L}$ such that for all interventions $\mathbf{f}'_{\mathcal{O}}$ that do not depend on $\mathcal{J}$ there exists some $\mathbf{g}'_{\mathcal{O}} : \mathbf{X}_{\mathcal{O}} \times \mathbf{E}_{\mathcal{J}} \to \mathbf{X}_{\mathcal{O}}$ that does not depend on $\mathcal{J}$ such that

$$(\mathbb{P}_{\mathcal{M}^{I(\mathcal{O})}, \mathbf{f}'_{\mathcal{O}}})^*_{\mathcal{L}} = \mathbb{P}_{(\mathcal{M}^{*_{\mathcal{L}}}, I(\mathcal{O}), \mathbf{g}'_{\mathcal{O}})}$$

We claim that given any SCNM $\mathcal{M}$ containing endogenous variables X_h, X_w and X_b representing height, weight and BMI there should be a primitive model $\mathcal{M}^p$ of $\mathcal{M}$ with respect to $\{p\}$.

Lemma 2.17 (Primitive models). $\mathcal{M}^p$ *is a primitive model of $\mathcal{M}$ with respect to $\mathcal{L} \subset \mathcal{I}$ iff $S(\pi_{\mathcal{O}} \mathbf{f}_{\mathcal{I}}) \overset{a.s.}{=} S(\mathbf{f}_{\mathcal{O}}^p)$ for $\mathcal{O} := \mathcal{I} \setminus \mathcal{L}$ and for all $\mathbf{X} \in \mathbf{X}_{\mathcal{I}}$, $\mathbf{g}$*

However, as Theroem 2.19 shows, if an SCNM with height, weight and BMI can be derived from an SCNM containing just height and weight then there are no valid hard interventions on BMI.

Definition 2.18 (Derived model). Given a SCNM $\mathcal{M} := \langle \mathcal{I}, \mathcal{J}, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{\mathcal{J}}, \mathbf{f}_{\mathcal{I}}, \mathbf{g}_{\mathcal{I}}, \mathbb{P}_{\mathcal{E}}, \boldsymbol{E}_{\mathcal{J}} \rangle$, say $\mathcal{M}' = \langle \mathcal{I}', \mathcal{J}, \mathbf{X}_{\mathcal{I}'}, \mathbf{E}_{\mathcal{J}}, \mathbf{f}'_{\mathcal{I}'}, \mathbf{g}'_{\mathcal{I}'}, \mathbb{P}_{\mathcal{E}}, \boldsymbol{E}_{\mathcal{J}} \rangle$ is *derived* from $\mathcal{M}$ if there exists some additional index/variable/relationships $i' \notin \mathcal{I}, X_{i'}$ such that

$$\mathcal{I}' = \mathcal{I} \cup \{i'\} \tag{57}$$
$$\mathbf{X}_{\mathcal{I}'} = \mathbf{X}_{\mathcal{I}} \cup X_{i'} \tag{58}$$

and, defining $\pi_{\mathcal{I}' \setminus i'} : \mathbf{X}_{\mathcal{I}'} \to \mathbf{X}_{\mathcal{I}}$ as the projection map that "forgets" $\mathsf{X}_{i'}$, for any $\mathbf{e} \in \mathbf{E}_{\mathcal{J}}$ we have

$$\mathbf{X}' = \mathbf{f}'_{\mathcal{I}'}(\mathbf{X}', \mathbf{e}) \tag{59}$$
$$\text{and } \mathbf{X}' = \mathbf{g}'_{\mathcal{I}'}(\mathbf{X}') \implies \pi_{\mathcal{I}' \setminus i'}(\mathbf{X}') \qquad = \mathbf{f}_{\mathcal{I}}(\pi_{\mathcal{I}' \setminus i'}(\mathbf{X}'), \mathbf{e}) \tag{60}$$
$$\text{and } \pi_{\mathcal{I}' \setminus i'}(\mathbf{X}') = \mathbf{g}'_{\mathcal{I}'}(\pi_{\mathcal{I}' \setminus i'}(\mathbf{X}')) \tag{61}$$

Theorem 2.19 (Interventions and necessary relationships don't mix). *If $\mathcal{M}'$ is derived from $\mathcal{M}$ with the additional elements $i', X_{i'}, f_{i'}, g_{i'}$ and both $\mathcal{M}$ and $\mathcal{M}'$ are uniquely solvable and $\mathbb{P}_{\mathcal{X}' \otimes \mathcal{E}}(\mathsf{X}_{i'})$ is not single valued then no hard interventions on $\mathsf{X}_{i'}$ are possible.*

Proof. Because $\mathcal{M}$ is uniquely solvable, for $\mathbb{P}_{\mathcal{E}}$ almost every $\mathbf{e}$ there is a unique $\mathbf{X}^e$ such that

$$\mathbf{X}^e = \mathbf{f}_{\mathcal{I}}(\mathbf{X}^e, \mathbf{e}) \tag{62}$$

$$\mathbf{X}^e = \mathbf{g}_{\mathcal{I}}(\mathbf{X}^e) \tag{63}$$

Because $\mathcal{M}'$ is also uniquely solvable, for $\mathbb{P}_{\mathcal{E}}$ almost every $\mathbf{e}$ we have $\mathbf{X}'^e \in \mathbf{X}_{\mathcal{I}'}$ such that $\pi_{\mathcal{I}' \setminus i'}(\mathbf{X}')'^e = \mathbf{X}^e$ and

$$x_{i'}'^e = \mathbf{g}_{i'}(\mathbf{X}'^e) \tag{64}$$

Because $\mathbb{P}_{\mathcal{X}' \otimes \mathcal{E}}(\mathsf{X}_{i'})$ is not single valued there are non-null sets $A, B \in \mathcal{E}$ such that $e_a \in A$, $e_b \in B$ implies

$$\mathbf{g}_{i'}(\mathbf{X}'^{e_a}) \neq \mathbf{g}_{i'}(\mathbf{X}'^{e_b}) \tag{65}$$

Therefore there exists no $a \in X_{i'}$ that can simultaneously satisfy 64 for almost every $\mathbf{e}$. However, any hard intervention $\mathcal{M}'^{,do(\mathsf{X}_{i'}=a)}$ requires such an a in order to be solvable. $\qquad \square$

Corollary 2.20. *Either there are no hard interventions defined on BMI or there is no SCNM containing height and weight with a unique solution from which an SCNM containing height, weight and BMI can be derived.*

> I can formalise the following, but I'm just writing it out so I can get to the end for now

The problem posed by Theorem 2.19 can be circumvented to some extent by joint interventions. Consider the variables X_1 and X_2 where $\mathsf{X}_1 = -\mathsf{X}_2$ necessarily. While Theorem 2.19 disallows interventions on X_2 individually (supposing we can obtain a unique model featuring only X_1), it does not disallow interventions that jointly set X_1 and X_2 to permissible values. In this case, this is unproblematic as the only joint intervention that sets X_1 to 1 must also set X_2 to -1.

If we have non-invertible necessary relationships such as $\mathsf{X}_1 = \mathsf{X}_2 + \mathsf{X}_3$, however, there are now *multiple* joint interventions on X_1 that can be performed. I regard this as the most plausible solution to the difficulties raised so far: for variables that are in non-invertible necessary relationships, there is a set of operations associated with the "intervention" that sets $\mathsf{X}_1 = 1$.

However, we still need to make sure the interventions that we have supposed comprise the operations associated with setting $\mathsf{X}_1 = 1$ exist themselves. It is sufficient that the SCNM with X_1 is derived from a higher order *uniquely solvable SCM* with X_2 and X_3 only .

> because interventions are defined in uniquely solvable SCMs and derivation preserves interventions on the old variables

If any variables are included in a causal model that are necessarily related to other variables (and honestly, is there any variable that isn't?), it is not enough to suppose that the model being used is a marginalisation of some larger causal model. Rather, it must be obtained by derivation and marginalisation from some model that represents the basic interventions that are possible, which we call the *atomic model*.

Definition 2.21 (Atomic model)**.** Given an SCNM $\mathcal{M}$, the *atomic model* $\mathcal{M}_{\text{atom}}$ is a uniquely solvable SCM such that there exists a model $\mathcal{M}$ is derived from of $\mathcal{M}_{\text{atom}}$.

Definition 2.22 (Causal universality hypothesis)**.** There exists a uniquely solvable SCM $\mathcal{M}_{\text{atom}}$ which is the atomic model that correctly represents all decision problems

3 Definitions and key notation

We use three notations for working with probability theory. The "elementary" notation makes use of regular symbolic conventions (functions, products, sums, integrals, unions etc.) along with the expectation operator $\mathbb{E}$. This is the most flexible notation which comes at the cost of being verbose and difficult to read. Secondly, we use a semi-formal string diagram notation extending the formal diagram notation for symmetric monoidal categories Selinger (2010). Objects in this diagram refer to stochastic maps, and by interpreting diagrams as symbols we can, in theory, be just as flexible as the purely symbolic approach. However, we avoid complex mixtures of symbols and diagrams elements, and fall back to symbolic representations if it is called for. Finally, we use a matrix-vector product convention that isn't particularly expressive but can compactly express some common operations.

3.1 Standard Symbols

Symbol	Meaning
$[n]$	The natural numbers $\{1, ..., n\}$
$f : a \mapsto b$	Function definition, equivalent to $f(a) := b$
Dots appearing in function arguments: $f(\cdot, \cdot, z)$	The "curried" function $(x, y) \mapsto f(x, y, z)$
Capital letters: A, B, X	sets
Script letters: $\mathcal{A}, \mathcal{B}, \mathcal{X}$	σ-algebras on the sets A, B, X respectively
Script $\mathcal{G}$	A directed acyclic graph made up of nodes V and edg
Greek letters μ, ξ, γ	Probability measures
δ_x	The Dirac delta measure: $\delta_x(A) = 1$ if $x \in A$ and 0 oth
Capital delta: $\Delta(\mathcal{E})$	The set of all probability measures on $\mathcal{E}$
Bold capitals: $\mathbf{A}$	Markov kernel $\mathbf{A} : X \times \mathcal{Y} \to [0, 1]$ (stochastic maps
Subscripted bold capitals: $\mathbf{A}_x$	The probability measure given by the curried Markov ker
$A \to \Delta(\mathcal{B})$	Markov kernel signature, treated as equivalent to $A \times \mathcal{B}$
$\mathbf{A} : x \mapsto \nu$	Markov kernel definition, equivalent to $\mathbf{A}(x, B) = \nu(B)$ f
Sans serif capitals: A, X	Measurable functions; we will also call them random va
$\mathbf{F}_{\mathsf{X}}$	The Markov kernel associated with the function X: $\mathbf{F}_{\mathsf{X}} \equiv$
$\mathbf{N}_{\mathsf{A}|\mathsf{B}}$	The conditional probability (disintegration) of A given B
$\nu\mathbf{F}_{\mathsf{X}}$	The marginal distribution of X under ν

3.2 Probability Theory

Given a set A, a σ-algebra $\mathcal{A}$ is a collection of subsets of A where

- $A \in \mathcal{A}$ and $\emptyset \in \mathcal{A}$

- $B \in \mathcal{A} \implies B^C \in \mathcal{A}$

- $\mathcal{A}$ is closed under countable unions: For any countable collection $\{B_i | i \in Z \subset \mathbb{N}\}$ of elements of $\mathcal{A}$, $\cup_{i \in Z} B_i \in \mathcal{A}$

A measurable space $(A, \mathcal{A})$ is a set A along with a σ-algebra $\mathcal{A}$. Sometimes the sigma algebra will be left implicit, in which case A will just be introduced as a measurable space.

Common σ algebras For any A, $\{\emptyset, A\}$ is a σ-algebra. In particular, it is the only sigma algebra for any one element set $\{*\}$.

For countable A, the power set $\mathcal{P}(A)$ is known as the discrete σ-algebra.

Given A and a collection of subsets of $B \subset \mathcal{P}(A)$, $\sigma(B)$ is the smallest σ-algebra containing all the elements of B.

Let T be all the open subsets of $\mathbb{R}$. Then $\mathcal{B}(\mathbb{R}) := \sigma(T)$ is the *Borel σ-algebra* on the reals. This definition extends to an arbitrary topological space A with topology T.

A *standard measurable set* is a measurable set A that is isomorphic either to a discrete measurable space A or $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For any A that is a complete separable metric space, $(A, \mathcal{B}(A))$ is standard measurable.

Given a measurable space $(E, \mathcal{E})$, a map $\mu : \mathcal{E} \to [0, 1]$ is a *probability measure* if

- $\mu(E) = 1$, $\mu(\emptyset) = 0$

- Given countable collection $\{A_i\} \subset \mathcal{E}$, $\mu(\cup_i A_i) = \sum_i \mu(A_i)$

Write by $\Delta(\mathcal{E})$ the set of all probability measures on $\mathcal{E}$.

Given a second measurable space $(F, \mathcal{F})$, a *stochastic map* or *Markov kernel* is a map $\mathbf{M} : E \times \mathcal{F} \to [0, 1]$ such that

- The map $\mathbf{M}(\cdot ; A) : x \mapsto \mathbf{M}(x; A)$ is $\mathcal{E}$-measurable for all $A \in \mathcal{F}$

- The map $\mathbf{M}_x : A \mapsto \mathbf{M}(x; A)$ is a probability measure on F for all $x \in E$

Extending the subscript notation above, for $\mathbf{C} : X \times Y \to \Delta(\mathcal{Z})$ and $x \in X$ we will write $\mathbf{C}_x$ for the "curried" map $y \mapsto \mathbf{C}_{x,y}$.

The map $x \mapsto \mathbf{M}_x$ is of type $E \to \Delta(\mathcal{F})$. We will abuse notation somewhat to write $\mathbf{M} : E \to \Delta(\mathcal{F})$, which captures the intuition that a Markov kernel maps from elements of E to probability measures on $\mathcal{F}$. Note that we "reverse" this idea and consider Markov kernels to map from elements of $\mathcal{F}$ to measurable functions $E \to [0, 1]$, an interpretation found in Clerc et al. (2017), but (at this stage) we don't make use of this interpretation here.

Given an indiscrete measurable space $(\{*\}, \{\{*\}, \emptyset\})$, we identify Markov kernels $\mathbf{N} : \{*\} \to \Delta(\mathcal{E})$ with the probability measure $\mathbf{N}_*$. In addition, there is a unique Markov kernel $* : E \to \Delta(\{\{*\}, \emptyset\})$ given by $x \mapsto \delta_*$ for all $x \in E$ which we will call the "discard" map.

3.3 Product Notation

We can use a notation similar to the standard notation for matrix-vector products to represent operations with Markov kernels. Probability measures $\mu \in \Delta(\mathcal{X})$ can be read as row vectors, Markov kernels as matrices and measurable functions $\mathsf{T} : Y \to T$ as column vectors. Defining $\mathbf{M} : X \to \Delta(\mathcal{Y})$ and $\mathbf{N} : Y \to \Delta(\mathcal{Z})$, the measure-kernel product $\mu\mathbf{A}(G) := \int \mathbf{A}_x(G)d\mu(x)$ yields a probability measure $\mu\mathbf{A}$ on $\mathcal{Z}$, the kernel-kernel product $\mathbf{MN}(x; H) = \int_Y \mathbf{B}(y; H)d\mathbf{A}_x$ yields a kernel $\mathbf{MN} : X \to \Delta(\mathcal{Z})$ and the kernel-function product $\mathbf{A}\mathsf{T}(x) := \int_Y \mathsf{T}(y)d\mathbf{A}_x$ yields a measurable function $X \to T$. Kernel products are associative (Çinlar, 2011).

The tensor product $(\mathbf{M} \otimes \mathbf{N})(x, y; G, H) := \mathbf{M}(x; G)\mathbf{N}(y; H)$ yields a kernel $(\mathbf{M} \otimes \mathbf{N}) : X \times Y \to \Delta(\mathcal{Y} \otimes \mathcal{Z})$.

3.4 String Diagrams

Some constructions are unwieldly in product notation; for example, given $\mu \in \Delta(\mathcal{E})$ and $\mathbf{M} : E \to (\mathcal{F})$, it is not straightforward to construct a measure $\nu \in \Delta(\mathcal{E} \otimes \mathcal{F})$ that captures the "joint distribution" given by $A \times B \mapsto \int_A \mathbf{M}(x; B)d\mu$.

Such constructions can, however, be straightforwardly captured with string diagrams, a notation developed for category theoretic probability. Cho and Jacobs (2019) also provides an extensive introduction to the notation discussed here.

Some key ideas of string diagrams:

- Basic string diagrams can always be interpreted as a mixture of kernel-kernel products and tensor products of Markov kernels

    - Extended string diagrams can be interepreted as a mixture of kernel-kernel products, kernel-function products, tensor products of kernels and functions and scalar products

- String diagrams are the subject of a coherence theorem: taking a string diagram and applying a planar deformation yields a string diagram that represents the same kernel (Selinger, 2010). This also holds for a number of additional transformations detailed below

A kernel $\mathbf{M} : X \to \Delta(\mathcal{Y})$ is written as a box with input and output wires, probability measures $\mu \in \Delta(\mathcal{X})$ are written as triangles "closed on the left" and measurable functions (which are only elements of the "extended" notation) $\mathsf{T} : Y \to T$ as triangles "closed on the right". For this introduction we will label wires with the names of their corresponding spaces, but in practice we will usually name them with corresponding *random variables*, though additional care is required when using random variables as labels (see paragraph 3.4.3).

For $\mathbf{M} : X \to \Delta(\mathcal{Y})$, $\mu \in \Delta(\mathcal{X})$ and $f : X \to W$:

$$X -\boxed{\mathbf{M}}- Y \qquad \triangleleft\!\mu\!\vdash X \qquad X -\!\!\rhd\!f \tag{66}$$

Elementary operations   We can compose Markov kernels with appropriate spaces - the equivalent operation of the "matrix products" of product notation. Given $\mathbf{M} : X \to \Delta(\mathcal{Y})$ and $\mathbf{N} : Y \to \Delta(\mathcal{Z})$, we have

$$\mathbf{MN} := \ X -\boxed{\mathbf{M}}\!-\!\boxed{\mathbf{N}}- Z \tag{67}$$

Probability measures are distinguished in that that they only admit "right composition" while functions only admit "left composition". For $\mu \in \Delta(\mathcal{E})$, $h : F \to X$:

$$\mu\mathbf{M} := \triangleleft\!\mu\vdash\!-\boxed{\mathbf{M}}- Z \tag{68}$$

$$\mathbf{M}f := \ X -\boxed{\mathbf{M}}\!-\!\rhd\!f \tag{69}$$

A diagram that is closed on the right and the left is an expectation:

$$\mathbb{E}_{\mu\mathbf{M}}(f) = \mu\mathbf{M}f \tag{70}$$

$$:= \quad \langle\!\!\mu\!\!-\!\!\boxed{\mathbf{M}}\!\!-\!\!f\rangle \tag{71}$$

We can also combine Markov kernels using tensor products, which we represent with vertical juxtaposition. For $\mathbf{O} : Z \to \Delta(\mathcal{W})$:

$$\mathbf{M} \otimes \mathbf{N} := \begin{array}{c} X -\!\boxed{\mathbf{M}}\!- Y \\ Z -\!\boxed{\mathbf{O}}\!- W \end{array} \tag{72}$$

Product spaces can be represented either by two parallel wires or a single wire:

$$X \times Y \cong \mathrm{Id}_X \otimes \mathrm{Id}_Y := \begin{array}{c} X -\!\!\!- X \\ Y -\!\!\!- Y \end{array} \tag{73}$$

$$= \quad X \times Y -\!\!\!\!- X \times Y \tag{74}$$

Because a product space can be represented by parallel wires, a kernel $\mathbf{L} : X \to \Delta(\mathcal{Y} \otimes \mathcal{Z})$ can be written using either two parallel output wires or a single output wire:

$$X -\!\boxed{\mathbf{L}}\!\begin{array}{c} Y \\ Z \end{array} \tag{75}$$

$$\equiv \tag{76}$$

$$X -\!\boxed{\mathbf{L}}\!- Y \times Z \tag{77}$$

Probability measures, Markov kernels and functions One has to exercise special care when including functions in diagrammatic notation. While any diagram that includes only probability measures (triangles pointing to the left) and Markov kernels (rectangles) is automatically a Markov kernel itself, while diagrams that include functions (triangles pointing to the right) only represent Markov kernels if they are correctly normalised, which is not a property that can be checked just by looking at the shape of the diagram.

Markov kernels with special notation A number of Markov kernels are given special notation distinct from the generic "box" representation above. These special representations facilitate intuitive graphical interpretations.

The identity kernel $\mathbf{Id} : X \to \Delta(X)$ maps a point x to the measure δ_x that places all mass on the same point:

$$\mathbf{Id}_x : x \mapsto \delta_x \equiv \; X \longrightarrow X \tag{78}$$

The copy map $\curlyvee : X \to \Delta(\mathcal{X} \times \mathcal{X})$ maps a point x to two identical copies of x:

$$\curlyvee : x \mapsto \delta_{(x,x)} \equiv \; X \prec \begin{matrix} X \\ X \end{matrix} \tag{79}$$

The swap map $\sigma : X \times Y \to \Delta(\mathcal{Y} \otimes \mathcal{X})$ swaps its inputs:

$$\sigma := (x,y) \to \delta_{(y,x)} \equiv \; \begin{matrix} Y \\ X \end{matrix} \asymp \begin{matrix} X \\ Y \end{matrix} \tag{80}$$

The discard map $* : X \to \Delta(\{*\})$ maps every input to δ_*. Note that the only non-empty event in $\{\emptyset, \{*\}\}$ must have probability 1.

$$* : x \mapsto \delta_* \equiv \; X \longrightarrow * \tag{81}$$

Any measurable function $F \to X$ has an associated Markov kernel $F \to \Delta(\mathcal{X})$. The Markov kernel associated with a function is different to the function itself - while the product of a probability measure μ with a function f is an expectation μf (see Definition 71), the product of a probability measure with the associated Markov kernel is the pushforward measure $f_\# \mu$.

Definition 3.1 (Function induced kernel). Given a measurable function $g : F \to X$, define the function induced kernel $\mathbf{F}_g : F \to \Delta(\mathcal{X})$ to be the the Markov kernel $a \mapsto \delta_{g(a)}$ for all $a \in X$.

Definition 3.2 (Pushforward kernel). Given a kernel $\mathbf{M} : E \to \Delta(\mathcal{F})$ and a measurable function $g : F \to X$, the *pushforward kernel* $g_\# \mathbf{M} : E \to \Delta(\mathcal{X})$ is the kernel such that $g_\# \mathbf{M}(a; B) = \mathbf{M}(a; g^{-1}(B))$.

If E is the one element space $\{*\}$, then $\mathbf{M} : \{*\} \to \Delta(\mathcal{F})$ can be identified with the probability measure $\mathbf{M}_*$ and the pushforward kernel $g_\# \mathbf{M}$ identified with the pushforward measure $g_\# \mathbf{M}_*$, so pushforward kernels reduce to pushforward measures.

Lemma 3.3 (Pushforward kernels are functional kernel products). *Given a kernel $\mathbf{M} : E \to \Delta(\mathcal{F})$ and a measurable function $g : F \to X$, the pushforward $g_\# \mathbf{M} = \mathbf{M} \mathbf{F}_g$.*

Proof.

$$\mathbf{MF}_g(a; B) = \int_F \delta_{g(y)}(B) d\mathbf{M}_a(y) \tag{82}$$

$$= \int_F \delta_y(g^{-1}(B)) d\mathbf{M}_a(y) \tag{83}$$

$$= \int_{g^{-1}(B)} d\mathbf{M}_a(y) \tag{84}$$

$$= g_\# \mathbf{M}(a; B) \tag{85}$$

$\square$

3.4.1 Comparison of notations

We are in a position to compare the three introduced notations using a few examples. Given $\mu \in \Delta(X), \mathbf{A} : X \to \Delta(Y)$ and $A \in \mathcal{X}, B \in \mathcal{Y}$, the following correspondences hold, where we express the same object in elementary notation, product notation and string notation respectively:

$$\nu := A \times B \mapsto \int_A A(x; B) d\mu(x) \equiv \mu \curlyvee (\mathbf{Id}_X \otimes \mathbf{A}) \equiv \quad \tag{86}$$

Where the resulting object is a probability measure $\nu \in \Delta(\mathcal{X} \otimes \mathcal{Y})$. Note that the elementary notation requires a function definition here, while the product and string notations can represent the measure without explicitly addressing its action on various inputs and outputs. Cho and Jacobs (2019) calls this construction "integrating $\mathbf{A}$ with respect to μ".

Define the marginal $\nu_Y \in \Delta(\mathcal{Y}) : B \mapsto \nu(X \times B)$ for $B \in \mathcal{Y}$ and similarly for ν_X. We can then express the result of marginalising 86 over X in our three separate notations as follows:

$$\nu_Y(B) = \nu(X \times B) = \int_X A(x; B) d\mu(x) \tag{87}$$

$$\nu_Y = \mu \mathbf{A} = \mu \curlyvee (\mathbf{Id}_X \otimes \mathbf{A})(\circledast \otimes \mathbf{Id}_Y) \tag{88}$$

$$\nu_Y = \qquad = \qquad \tag{89}$$

The elementary notation 87 makes the relationship between ν_Y and ν explicit and, again, requires the action on each event to be defined. The product notation 88 is, in my view, the least transparent but also the most compact in the form $\mu \mathbf{A}$, and does not demand the explicit definition of how ν_Y treats every event. The graphical notation is the least compact in terms of space taken

up on the page, but unlike the product notation it shows a clear relationship to the graphical construction in86, and displays a clear graphical logic whereby marginalisation corresponds to "cutting off branches". Like product notation, it also allows for the definition of derived measures such as ν_Y without explicit definition of the handling of all events. It also features a much smaller collection of symbols than does elementary notation.

String diagrams often achieve a good balance between being ease of understanding at a glance and expressive power. On the downside, they can be time consuming to typeset, and formal reasoning with them takes some practice.

3.4.2   Working With String Diagrams

todo:

- Functional generalisation

- Conditioning

- Infinite copy map

- De Finetti's representation theorem

There are a relatively small number of manipulation rules that are useful for string diagrams. In addition, we will define graphically analogues of the standard notions of *conditional probability*, *conditioning*, and infinite sequences of exchangeable random variables.

Axioms of Symmetric Monoidal Categories   For the following, we either omit labels or label diagrams with their domain and codomain spaces, as we are discussing identities of kernels rather than identities of components of a condtional probability space. Recalling the unique Markov kernels defined above, the following equivalences, known as the *commutative comonoid axioms*, hold among string diagrams:

$$\tag{90}$$

$$\tag{91}$$

$$\tag{92}$$

The discard map $*$ can "fall through" any Markov kernel:

$$—\boxed{\mathbf{A}}—* \quad = \quad —* \tag{93}$$

Combining 91 and 93 we can derive the following: integrating $\mathbf{A} : X \to \Delta(\mathcal{Y})$ with respect to $\mu \in \Delta(\mathcal{X})$ and then discarding the output of $\mathbf{A}$ leaves us with μ:

$$\tag{94}$$

In elementary notation, this is equivalent to the fact that, for all $B \in \mathcal{X}$, $\int_B \mathbf{A}(x; B) d\mu(x) = \mu(B)$.

The following additional properties hold for $*$ and $\curlyvee$:

$$X \times Y \longrightarrow * \quad = \quad \begin{matrix} X \longrightarrow * \\ Y \longrightarrow * \end{matrix} \tag{95}$$

$$\tag{96}$$

A key fact that *does not* hold in general is

$$\tag{97}$$

In fact, it holds only when $\mathbf{A}$ is a *deterministic* kernel.

Definition 3.4 (Deterministic Markov kernel)**.** A *deterministic* Markov kernel $\mathbf{A} : E \to \Delta(\mathcal{F})$ is a kernel such that $\mathbf{A}_x(B) \in \{0, 1\}$ for all $x \in E$, $B \in \mathcal{F}$.

Theorem 3.5 (Copy map commutes for deterministic kernels (Fong, 2013))**.** *Equation 97 holds iff $\mathbf{A}$ is deterministic.*

3.4.3   Random Variables

The summary of this section is:

- Random variables are usually defined as measurable functions on a *probability space*

- It's possible to define them as measurable functions on a *Markov kernel space* instead

- It is useful to labelling wires with random variable names instead of names of spaces

Probability theory is primarily concerned with the behaviour of *random variables*. This behaviour can be analysed via a collection of probability measures and Markov kernels representing joint, marginal and conditional distributions of random variables of interest. In the framework developed by Kolmogorov, this collection of joint, marginal and conditional distributions is modeled by a single underlying *probability space*, and random variables by measurable functions on the probability space.

We use the same approach here, with a couple of additions. We are interested in variables whose outcomes depend both on random processes and decisions. These variables are better modelled by a Markov kernels than probability measure - *given* a particular decision, they inherit a particular probability distribution. Thus, variables in our work are modeled by an underlying Markov kernel rather than a probability measure; we call this a *Markov kernel space*.

In addition to following standard conventions regarding the use of random variables, we can motivate their introduction with a graphical example. Suppose we have some $\mu \in \Delta(\mathcal{X} \otimes \mathcal{X})$, $\mathbf{K} : X \to \Delta(\mathcal{X})$ such that the following holds:

$$\mu \vDash \begin{matrix} X \\ X \end{matrix} \quad = \quad \mu \vdash * \multimap \boxed{\mathbf{K}} \vdash \begin{matrix} X \\ X \end{matrix} \tag{98}$$

This implies, roughly, that $\mathbf{K}$ is the probability of the lower wire of μ conditional on the upper (it is a *disintegration* of μ, defined later). However, it is very cumbersome to write 98 and define $\mathbf{K}$ in terms of the geometry of the diagrams. Instead, it would be nice to have a system where we can unambiguously assign *names* to wires:

$$\mu \vDash \begin{matrix} \mathsf{X}_1 \\ \mathsf{X}_2 \end{matrix} \tag{99}$$

Once the wires of μ have names, we can define a convention such that:

$$\mathsf{X}_1 \multimap \boxed{\overline{}_{\mathsf{Y}|\mathsf{X}}} \multimap \mathsf{Y}_2 \tag{100}$$

Is a Markov kernel $X \to \Delta(\mathcal{X})$ that satisfies Equation 98 when substituted for $\mathbf{K}$. We take wire names to stand for random variables on a *Markov kernel space*.

Definition 3.6 (Probability space, Markov kernel space)**.** A *probability space* $(\mathbb{P}, \Omega, \mathcal{F})$ is a probability measure $\mathbb{P}$, which we call the *ambient measure*, along with the *sample space* Ω and the *events* $\mathcal{F}$.

A *Markov kernel space* $(\mathbb{K}, \Omega, \mathcal{F}, D, \mathcal{D})$ is a Markov kernel $\mathbb{K} : D \to \Delta(\mathcal{D} \otimes \mathcal{F})$, called the *ambient kernel*, along with the sample space $(\Omega, \mathcal{F})$ and the domain $(D, \mathcal{D})$. We suppose that $\mathbb{K}$ is such that there exists a *fundamental kernel* $\mathbb{K}_0$ satisfying

$$\mathbb{K} := \quad \begin{array}{c}\boxed{\mathbb{K}_0}\end{array} \tag{101}$$

It is in general much more practical to work with $\mathbb{K}$ than $\mathbb{K}_0$.

> Is this sufficient to make kernel spaces a special case of conditional probability spaces?

Definition 3.7 (Random variable)**.** Given a sample space Ω and an domain D, a random variable X is a measurable function $\Omega \times D \to E$ for arbitrary measurable E.

Definition 3.8 (Domain variable)**.** Given a sample space Ω and an domain D, the *domain variable* $\mathsf{D} : \Omega \times D \to D$ is the random variable given by $\mathsf{D} : (x, d) \mapsto d$.

Unlike random variables on probability spaces, random variables on Markov kernel spaces do not in general have unique marginal distributions. An analogous operation of *marginalisation* can be defined, but the result is generally a Markov kernel.

Definition 3.9 (Coupled tensor product $\underline{\otimes}$)**.** Given two Markov kernels $\mathbf{M}$ and $\mathbf{N}$ or functions f and g with shared domain E, let $\mathbf{M} \underline{\otimes} \mathbf{N} := \curlyvee(\mathbf{M} \otimes \mathbf{N})$ and $f \underline{\otimes} g := \curlyvee(f \otimes g)$ where these expressions are interpreted using standard product notation. Graphically:

$$\mathbf{M} \underline{\otimes} \mathbf{N} := \quad E \; \begin{array}{c}\boxed{\mathbf{M}}\!-\!\mathsf{X} \\ \boxed{\mathbf{N}}\!-\!\mathsf{Y}\end{array} \tag{102}$$

$$f \underline{\otimes} g := \quad E \; \begin{array}{c}\triangleright f \\ \triangleright g\end{array} \tag{103}$$

The operation denoted by $\underline{\otimes}$ is associative (Lemma 3.10), so we can without ambiguity write $f \underline{\otimes} g \underline{\otimes} h = (f \underline{\otimes} g) \underline{\otimes} h = f \underline{\otimes} (g \underline{\otimes} h)$ for finite groups of functions or Markov kernels sharing a domain.

The notation $\underline{\otimes}_{i \in [N]} f_i$ is taken to mean $f_1 \underline{\otimes} f_2 \underline{\otimes} ... \underline{\otimes} f_N$.

Lemma 3.10 ($\underline{\otimes}$ is associative)**.** *For Markov kernels* $\mathbf{L} : E \to \delta(\mathcal{F})$, $\mathbf{M} : E \to \delta(\mathcal{G})$ *and* $\mathbf{N} : E \to \delta(\mathcal{H})$, $(\mathbf{L} \underline{\otimes} \mathbf{M}) \underline{\otimes} \mathbf{N} = \mathbf{L} \underline{\otimes} (\mathbf{M} \underline{\otimes} \mathbf{N})$.

Proof.

$$\mathbf{L}\underline{\otimes}(\mathbf{M}\underline{\otimes}\mathbf{N}) = \qquad\qquad (104)$$

$$= \qquad\qquad (105)$$

$$= (\mathbf{L}\underline{\otimes}\mathbf{M})\underline{\otimes}\mathbf{N} \qquad (106)$$

This follows directly from Equation 90. $\qquad\square$

Definition 3.11 (Marginal distribution, marginal kernel)**.** Given $\mathbb{P} \in \Delta(\mathcal{F})$, random variable $\mathsf{X} : \Omega \to G$ the *marginal distribution* of X $\mathbb{P}_{\mathsf{X}} \in \Delta(\mathcal{G})$ of X is the product measure $\mathbb{P}\mathbf{F}_{\mathsf{X}}$.

See Lemma 3.3 for the proof that this matches the usual definition of marginal distribution.

Given $\mathbb{K} : D \to \Delta(\mathcal{F})$ and random variable $\mathsf{X} : \Omega \to G$, the *marginal kernel* is $\mathbb{K}_{\mathsf{X}|\mathsf{D}} := \mathbb{K}\mathbf{F}_{\mathsf{X}}$.

Definition 3.12 (Joint distribution, joint kernel)**.** Given $\mathbb{P} \in \Delta(\mathcal{F})$, $\mathsf{X} : \Omega \to G$ and $\mathsf{Y} : \Omega \to H$, the *joint distribution* $\mathbb{P}_{\mathsf{XY}} \in \Delta(\mathcal{G}\otimes\mathcal{H})$ of X and Y is the marginal distribution of $\mathsf{X}\underline{\otimes}\mathsf{Y}$.

This is identical to the definition in Çinlar (2011) if we note that the random variable $(\mathsf{X}, \mathsf{Y}) : \omega \mapsto (\mathsf{X}(\omega), \mathsf{Y}(\omega))$ (Çinlar's definition) is precisely the same thing as $\mathsf{X}\underline{\otimes}\mathsf{Y}$.

Analogously, the joint kernel $\mathbb{K}_{\mathsf{XY}|\mathsf{D}}$ is the product $\mathbb{K}\mathbf{F}_{\mathsf{X}\underline{\otimes}\mathsf{Y}}$.

Joint distributions and kernels have a nice visual representation, as a result of Lemma 3.13 which follows.

Lemma 3.13 (Dual representation of coupled products of functions)**.** *Given two functions, the kernel associated with their coupled product is equal to the coupled product of the kernels associated with each function.*

Given $\mathsf{X} : \Omega \to G$ and $\mathsf{Y} : \Omega \to H$, $\mathbf{F}_{\mathsf{X}\underline{\otimes}\mathsf{Y}} = \mathbf{F}_{\mathsf{X}}\underline{\otimes}\mathbf{F}_{\mathsf{Y}}$

Proof. For $a \in \Omega$, $B \in \mathcal{G}$, $C \in \mathcal{H}$,

$$\mathbf{F}_{\mathsf{X}\underline{\otimes}\mathsf{Y}}(a; B \times C) = \delta_{\mathsf{X}(a),\mathsf{Y}(a)}(B \times C) \qquad (107)$$

$$= \delta_{\mathsf{X}(a)}(B)\delta_{\mathsf{Y}(a)}(C) \qquad (108)$$

$$= (\delta_{\mathsf{X}(a)} \otimes \delta_{\mathsf{Y}(a)})(B \times C) \qquad (109)$$

$$= \mathbf{F}_{\mathsf{X}}\underline{\otimes}\mathbf{F}_{\mathsf{Y}} \qquad (110)$$

Equality follows from the monotone class theorem. $\qquad\square$

29

Corollary 3.14. *Given a Markov kernel space* $(\Omega, D, \mathbb{K})$ *and random variables* $\mathsf{X} : \Omega \times D \to X$, $\mathsf{Y} : \Omega \times D \to Y$, *the following holds:*

$$D -\boxed{\mathbb{K}_{\mathsf{XY|D}}}- \begin{matrix} X \\ Y \end{matrix} \quad = \quad D -\boxed{\mathbb{K}} - \left(\begin{matrix} \boxed{\mathbf{F_X}} - X \\ \boxed{\mathbf{F_Y}} - Y \end{matrix} \right. \tag{111}$$

We will now define wire labels for "output" wires.

Definition 3.15 (Wire labels - joint kernels)**.** Suppose we have a Markov kernel space $(\Omega, D, \mathbb{K})$ and random variables $\mathsf{X} : \Omega \times D \to X$, $\mathsf{Y} : \Omega \times D \to Y$, and a Markov kernel $\mathbf{L} : D \to \Delta(\mathcal{X} \times \mathcal{Y})$.

Relative to $(\Omega, D, \mathbb{K})$, the wires terminating on a free end on the right of a diagram of $\mathbf{L}$ may be labelled with X and Y as follows:

$$D -\boxed{\mathbf{L}}- \begin{matrix} \mathsf{X} \\ \mathsf{Y} \end{matrix} \tag{112}$$

iff

$$\mathbf{L} = \mathbb{K}_{\mathsf{XY|D}} \tag{113}$$

and

$$D -\boxed{\mathbf{L}} \begin{matrix} \mathsf{X} \\ {}_{*} \end{matrix} \quad = \mathbb{K}_{\mathsf{X|D}} \tag{114}$$

and

$$D -\boxed{\mathbf{L}} \begin{matrix} {}^{*} \\ \mathsf{Y} \end{matrix} \quad = \mathbb{K}_{\mathsf{Y|D}} \tag{115}$$

The second and third conditions are nontrivial: suppose X takes values in some product space $Range(\mathsf{X}) = W \times Z$, and Y takes values in Y. Then we could have $\mathbf{L} = \mathbb{K}_{\mathsf{XY|D}}$ and draw the diagram

$$D -\boxed{\mathbf{L}}- \begin{matrix} W \\ Z \end{matrix} \times Y \tag{116}$$

For *this* diagram, properties 114 and 115 do not hold, even though 113 does.

> I need to prove that if 113 holds and the spaces match the codomains of the random variables, then labels can be assigned

Having defined output wire labels, I will now proceed to use them without special colouring.

Definition 3.16 (Disintegration)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$, random variables X and Y and joint probability measure $\mathbb{P}_{\mathsf{XY}} \in \Delta(\mathcal{E} \otimes \mathcal{F})$, we say that $\mathbf{M} : E \to \Delta(\mathcal{F})$ is a Y *on* X *disintegration* of μ iff

$$\tag{117}$$

$\mathbf{M}$ is a version of $\mathbb{P}_{\mathsf{Y}|\mathsf{X}}$, "the probability of Y given X". Let $\mathbb{P}_{\{\mathsf{Y}|\mathsf{X}\}}$ be the set of all kernels that satisfy 117 and $\mathbb{P}_{\mathsf{Y}|\mathsf{X}}$ an arbitrary member of $\mathbb{P}_{\mathsf{Y}|\mathsf{X}}$.

Given a Markov kernel space $(\mathbb{K}, \Omega, D)$ and random variables $\mathsf{X} : \Omega \times D \to X$, $\mathsf{Y} : \Omega \times D \to Y$, $\mathbf{M} : D \times E \to \Delta(\mathcal{F})$ is a Y *on* $\mathsf{D}\mathsf{X}$ *disintegration* of $\mathbb{K}_{\mathsf{YX}|\mathsf{D}}$ iff

$$\tag{118}$$

Write $\mathbb{K}_{\{\mathsf{Y}|\mathsf{XD}\}}$ for the set of kernels satisfying 118 and $\mathbb{K}_{\mathsf{Y}|\mathsf{XD}}$ for an arbitrary member of $\mathbb{K}_{\{\mathsf{Y}|\mathsf{XD}\}}$.

Note that for any variable $\mathsf{X} : \Omega \times D \to X$ and the domain variable $\mathsf{D} : \Omega \times D \to D$ we have by definition of $\mathbb{K}$:

$$\tag{119}$$

$$\tag{120}$$

$$\tag{121}$$

$$\tag{122}$$

$$\tag{123}$$

That is, any joint kernel including the variable D can be drawn such that the wire labeled D is copied from the input wire. Conversely, if we have a joint kernel $\mathbb{K}_{\mathsf{X}|\mathsf{D}}$ and add a wire copied from the input, we now have $\mathbb{K}_{\mathsf{XD}|\mathsf{D}}$. We use this insight to give names to input wires: if copying an input wire to the output allows us to label the *output* wire with X, then the *input* wire will also be labeled X.

31

Definition 3.17 (Wire labels - disintegrations). Given a conditional probability space with ambient kernel $\mathcal{K} : D \to \Delta(\mathcal{F})$ (or a probability space with measure $\mathbb{P}$),

Note that $\mathbb{P}^*$ is simply $\mathbb{P}$ for a probability space

Recall that D is the global conditioning variable. Given two collections of random variables $c_1 = [\mathsf{X}_1, \mathsf{X}_2, ...]$ and $c_2 = [\mathsf{Y}_1, \mathsf{Y}_2]$, we adopt the convention that any diagram with the input wires labeled with c_1 and the output wires labeled with c_2 is an element of $\mathcal{K}^*_{\mathsf{Y}_1\mathsf{Y}_2...|\mathsf{X}_1\mathsf{X}_2...}$.

That is, by this convention, the diagram

$$\mathsf{X} \atop \mathsf{D} \ \boxed{\mathbf{M}} \ \mathsf{Y} \tag{124}$$

implies that $\mathbf{M} \in \mathcal{K}_{\mathsf{Y}|\mathsf{XD}}$. Note further that by Theorem 3.19, we can rely on the existence of disintegrations such as $\mathbf{M}$ that are conditional on the global conditioning variable D provided we have countable D and standard measurable $(Y, \mathcal{Y})$.

If we have some version $\mathbf{M}$ of $\mathcal{K}_{\mathsf{Y}|\mathsf{XD}}$ that does not depend on the value of D - i.e. $\mathbf{M}_{(x,d)} = \mathbf{M}_{(x,d')}$ for all $x \in X$, $d, d' \in D$, then there exists some $\mathbf{M}'$ such that:

$$\mathsf{X} \atop \mathsf{D} \ \boxed{\mathbf{M}} \ \mathsf{Y} \quad = \quad \begin{matrix} \mathsf{X} \ \boxed{\mathbf{M}'} \ \mathsf{Y} \\ \mathsf{D} \ \text{———} \ * \end{matrix} \tag{125}$$

Under these circumstances, we will abuse notation to say $\mathbf{M}' = \mathcal{K}_{\mathsf{Y}|\mathsf{X}}$.

We can't expect Equation 125 to hold in an arbitrary conditional probability spaces. For a very simple example, take $\mathcal{K} : \{0, 1\} \to \Delta(\{0, 1\})$ where $\mathcal{K}_0 = \mathcal{K}_1 = \text{Bernoulli}(0.5)$, and let $\mathsf{X} : (x, d) \mapsto x$ - i.e. the random variable projecting the output of $\mathcal{K}$. Then there is no disintegration $\mathcal{K}_{\mathsf{D}|\mathsf{X}}$ - we can't recover the input D from X.

Under some (strong) regularity conditions, disintegrations of conditional probability spaces do exist.

Theorem 3.18 (Disintegration existence - probability space). *Given a probability measure $\mu \in \Delta(\mathcal{E} \otimes \mathcal{F})$, if $(F, \mathcal{F})$ is standard then a disintegration $\mathbf{K} : E \to \Delta(\mathcal{F})$ exists (Çinlar, 2011).*

Theorem 3.19 (Disintegration existence - conditional probability space). *Given a kernel $\mathbf{L} : D \to \Delta(\mathcal{E} \otimes \mathcal{F})$, define $\mathbf{L}^*$:*

$$\boxed{\mathbf{L}} \tag{126}$$

If D is countable and $(F, \mathcal{F})$ is standard, then there is a disintegration $\mathbf{M}$: $D \times E \to \Delta(\mathcal{F})$ of $\mathbf{L}^*$.

Proof. By Theorem 3.18, for each $d \in D$ we have a disintegration $\mathbf{K}^{(d)} : E \to \Delta(\mathcal{F})$ of $\mathbf{L}_d$. Define $\mathbf{M} : D \times E \to \Delta(\mathcal{F})$ by $\mathbf{M}(d, e; A) = \mathbf{K}^{(d)}(e; A)$ for $d \in D$, $e \in E$, $A \in \mathcal{F}$. Clearly $\mathbf{M}_{(d,e)}$ is a probability measure. Furthermore, for $B \in \mathcal{B}(\mathbb{R})$, $\mathbf{M}^{-1}(\cdot; A)(B) = \cup_{d \in D}\{d\} \times \mathbf{K}^{(d)-1}(\cdot; A)(B)$, which is a countable union of measurable sets and therefore measurable. $\qquad\square$

As an aside, Hájek (2003) pointed out that in general there are many Markov kernels that satisfy the definition of conditional probability for a given probability measure and random variables. While it is interesting that for a given Markov kernel space $(\mathbb{K}, \Omega, \mathcal{F}, D, \mathcal{D})$ there is in general no probability measure on $\Omega \times D$ such that $\mathbb{K}$ is uniquely defined as a disintegration of μ. By limiting D to countable sets, we avoid this possibility, and limit ourselves to Markov kernel spaces that can be uniquely defined as disintegraitons.

From here on out, we will assume whether explicitly stated or not that any global conditioning space is countable and any other measureable space is standard, guaranteeing the existence of disintegrations.

In general, we don't want to spent time explicitly setting up conditional probability spaces. Rather, we will specify key marginals and disintegrations from which a conditional probability space can be constructed - call these marginals and conditional "components". Clearly we cannot build a conditional probability space from two kernels that represent the same component but disagree with each other on a non-negligible set. Also, in general, for an arbitrary collection of components there may be many ambient kernels from which we can extract these components. There is no particular problem if we have multiple ambient kernels over undefined random variable; if we are only interested in X then the possibility of many joint kernels over X and Y is no cause for concern. We do, however, want to avoid ambient kernels supporting non-negligibly distinct marginals or disintegrations over the random variables that have been defined.

Example 3.20 (Implicit conditional probability space)**.** Suppose we have labeled Markov kernels

$$\mathsf{D} -\boxed{\mathbf{L}}- \mathsf{X} \qquad \mathsf{X} -\boxed{\mathbf{M}}- \mathsf{Y} \qquad\qquad (127)$$

We want to define a conditional probability space $(\mathcal{K}, \Omega, D)$ supporting random variables D, X and Y yielding the above kernels as the relevant marginals and disintegrations. Strictly:

- $\mathbf{L} = \mathcal{K}_{\mathsf{X}|\mathsf{D}}$

- $\mathbf{M} \otimes {}^{*}_{D} \in \mathcal{K}_{\mathsf{Y}|\mathsf{X}D}$ ("informally", $\mathbf{M} \in \mathcal{K}_{\mathsf{Y}|\mathsf{X}}$)

Take $\Omega = W \times X \times Y \times Z$ and define $\mathcal{K}$ such that

$$\mathcal{K}^* = \quad \text{(diagram: } D \to \boxed{\mathbf{L}} \to \boxed{\mathbf{M}} \to Y,\ X,\ D\text{)}$$ (128)

Where $\mathcal{K}^*$ is the copy map composed with $\mathcal{K}$ as in previous definitions. $\mathcal{K}$ is the unique Markov kernel $D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ supporting the two criteria above, assuming finite D and standard measurable X, Y.

Proof. By assumption, for any suitable $\mathcal{K}' : D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ we have

$$D \to \boxed{\mathcal{K}'}_* X \quad = \quad D \to \boxed{\mathbf{L}} \to X$$ (129)

and by the fact that $\mathbf{M} \otimes \mathord{\dagger}_D$ is by assumption a disintegration of $\mathcal{K}'^*$:

$$\text{(diagram left)} \quad = \quad \text{(diagram right)}$$ (130)

$$= \quad \text{(diagram: } D \to \boxed{\mathbf{L}} \to \boxed{\mathbf{M}} \to Y,\ X,\ D\text{)}$$ (131)

Finally, if $\mathcal{K}^* = \mathcal{K}'^*$, then at least if D is countable we must have $\mathcal{K} = \mathcal{K}'$ as they must agree on all points in D. $\square$

This example was chosen to illustrate a peculiarity of our notation of conditional probability spaces. Consider a problem that appears similar: find an ambient measure $\mathbb{P}$ decomposing into the following marginal and conditionals:

$$\boxed{\mu} \to D \qquad D \to \boxed{\mathbf{L}} \to X \qquad X \to \boxed{\mathbf{M}} \to Y$$ (132)

Here there are many choices of $\mathbb{P}$ that satisfy our conditions arising from different choices of $\mathbb{P}_{Y|XD}$. This is not possible in the conditional probability space because $\mathcal{K}_{Y|X}$ only exists if $\mathcal{K}_{Y|XD}$ is independent of D. That is, in a conditional probability space every disintegration is conditional on D, but we may not explicitly write this if it does not actually depend on D.

A sufficient condition for the construction of a unique ambient kernel from a collection of components $\{C_1, ...C_n\}$ is if there is some ordering of components $\{i_1, i_2, ...i_n\}$ such that the input labels of $C_{i_{k+1}}$ is the union of the inputs and outputs of $C_{i_1}, ..., C_{i_j}$. This can be shown by repeated application of Theorem 3.19.

In general, diagram labels are "well behaved" with regard to the application of any of the special Markov kernels: identities 78, swaps 80, discards 81 and copies 79 as well as with respect to the coherence theorem of the CD category. They are not "well behaved" with respect to composition.

Lemma 3.21 (Diagrammatic consequences of labels)**.** *Fix some conditional probability space* $(\mathcal{K}, \Omega, D)$ *and random variables* X, Y, Z *taking values in arbitrary spaces.* Sat : *indicates that a labeled diagram satisfies definitions 3.15 and 3.17 with respect to* $(\mathcal{K}, \Omega, D)$ *and* X, Y, Z. *The following always holds:*

$$\text{Sat} : \mathsf{X} - \mathsf{X} \tag{133}$$

and the following implications hold:

$$\text{Sat} : \mathsf{Z} - \boxed{K} \begin{smallmatrix} \mathsf{X} \\ \mathsf{Y} \end{smallmatrix} \implies \text{Sat} : \mathsf{Z} - \boxed{K} \begin{smallmatrix} \mathsf{X} \\ * \end{smallmatrix} \tag{134}$$

$$\text{Sat} : \mathsf{Z} - \boxed{K} \begin{smallmatrix} \mathsf{X} \\ \mathsf{Y} \end{smallmatrix} \implies \text{Sat} : \mathsf{Z} - \boxed{K} \begin{smallmatrix} \mathsf{Y} \\ \mathsf{X} \end{smallmatrix} \tag{135}$$

$$\text{Sat} : \mathsf{Z} - \boxed{L} - \mathsf{X} \implies \text{Sat} : \mathsf{Z} - \boxed{L} - \begin{smallmatrix} \mathsf{X} \\ \mathsf{X} \end{smallmatrix} \tag{136}$$

$$\text{Sat} : \mathsf{Z} - \boxed{\mathbf{K}} - \mathsf{Y} \implies \text{Sat} : \mathsf{Z} - \overset{\mathsf{Z}}{\underset{\boxed{\mathbf{K}} - \mathsf{Y}}{\frown}} \tag{137}$$

Proof. • Id_X is a version of $\mathbb{P}_{\mathsf{X}|\mathsf{X}}$ for all $\mathbb{P}$; $\mathbb{P}_{\mathsf{X}}\text{Id}_X = \mathbb{P}_{\mathsf{X}}$

• $\mathbf{K}\text{Id} \otimes *)(w; A) = \int_{X \times Y} \delta_x(A)\mathbb{1}_Y(y)d\mathbf{K}_w(x,y) = \mathbf{K}_w(A \times Y) = \mathbb{P}_{\mathsf{X}|\mathsf{Z}}(w; A)$

• $\int_{X \times Y} \delta_{\text{swap}(x,y)}(A \times B)d\mathbf{K}_w(x,y) = \mathbb{P}_{\mathsf{YX}|\mathsf{Z}}(w; A \times B)$

• $\mathbf{K}\curlyvee(w; A \times B) = \int_X \delta_{x,x}(A \times B)d\mathbf{K}_w(x) = \mathbb{P}_{\mathsf{XX}|\mathsf{Z}}(w; A \times B)$
137: Suppose $\mathbf{K}$ is a version of $\mathbb{P}_{\mathsf{Y}|\mathsf{Z}}$. Then

$$\mathbb{P}_{\mathsf{ZY}} = \overset{\triangleleft \boxed{\mathbb{P}_{\mathsf{Z}}}}{} \frown \boxed{\mathbf{K}} - \begin{smallmatrix} \mathsf{Z} \\ \mathsf{Y} \end{smallmatrix} \tag{138}$$

$$\mathbb{P}_{\mathsf{ZZY}} = \overset{\triangleleft \boxed{\mathbb{P}_{\mathsf{Z}}}}{} \frown \boxed{\mathbf{K}} - \begin{smallmatrix} \mathsf{Z} \\ \mathsf{Z} \\ \mathsf{Y} \end{smallmatrix} \tag{139}$$

$$= \overset{\triangleleft \boxed{\mathbb{P}_{\mathsf{Z}}}}{} \frown \boxed{\mathbf{K}} - \begin{smallmatrix} \mathsf{Z} \\ \mathsf{Z} \\ \mathsf{Y} \end{smallmatrix} \tag{140}$$

Therefore $\curlyvee(\text{Id}_X \otimes \mathbf{K})$ is a version of $\mathbb{P}_{\mathsf{ZY}|\mathsf{Z}}$ by **??** ∎

35

The following property, on the other hand, does *not* generally hold:

$$\text{Sat}: \mathsf{Z} \!-\!\boxed{\mathsf{K}}\!-\! \mathsf{Y} \ , \ \mathsf{Y} \!-\!\boxed{\mathsf{L}}\!-\! \mathsf{X} \ \implies \ \text{Sat}: \mathsf{Z} \!-\!\boxed{\mathsf{K}}\!-\!\boxed{\mathsf{L}}\!-\! \mathsf{X} \qquad (141)$$

Consider some ambient measure $\mathbb{P}$ with $\mathsf{Z} = \mathsf{X}$ and $\mathbb{P}_{\mathsf{Y}|\mathsf{X}} = x \mapsto \text{Bernouli}(0.5)$ for all $z \in Z$. Then $\mathbb{P}_{\mathsf{Z}|\mathsf{Y}} = y \mapsto \mathbb{P}_{\mathsf{Z}}$, $\forall y \in Y$ and therefore $\mathbb{P}_{\mathsf{Y}|\mathsf{Z}}\mathbb{P}_{\mathsf{Z}|\mathsf{Y}} = x \mapsto \mathbb{P}_{\mathsf{Z}}$ but $\mathbb{P}_{\mathsf{Z}|\mathsf{X}} = x \mapsto \delta_x \neq \mathbb{P}_{\mathsf{Y}|\mathsf{Z}}\mathbb{P}_{\mathsf{Z}|\mathsf{Y}}$.

References

Stephan Bongers, Jonas Peters, Bernhard Schölkopf, and Joris M. Mooij. Theoretical Aspects of Cyclic Structural Causal Models. *arXiv:1611.06221 [cs, stat]*, November 2016. URL http://arxiv.org/abs/1611.06221. arXiv: 1611.06221.

Erhan Çinlar. *Probability and Stochastics*. Springer, 2011.

Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, August 2019. ISSN 0960-1295, 1469-8072. doi: 10.1017/S0960129518000488.

Florence Clerc, Fredrik Dahlqvist, Vincent Danos, and Ilias Garnier. Pointless learning. *20th International Conference on Foundations of Software Science and Computation Structures (FoSSaCS 2017)*, March 2017. doi: 10.1007/978-3-662-54458-7_21. URL https://www.research.ed.ac.uk/portal/en/publications/pointless-learning(694fb610-69c5-469c-9793-825df4f8ddec).html.

A. P. Dawid. Influence Diagrams for Causal Modelling and Inference. *International Statistical Review*, 70(2):161–189, 2002. ISSN 1751-5823. doi: 10.1111/j.1751-5823.2002.tb00354.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2002.tb00354.x.

Brendan Fong. Causal Theories: A Categorical Perspective on Bayesian Networks. *arXiv:1301.6201 [math]*, January 2013. URL http://arxiv.org/abs/1301.6201. arXiv: 1301.6201.

M. A. Hernán and S. L. Taubman. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity*, 32(S3):S8–S14, August 2008. ISSN 1476-5497. doi: 10.1038/ijo.2008.82. URL https://www.nature.com/articles/ijo200882.

Alan Hájek. What Conditional Probability Could Not Be. *Synthese*, 137(3): 273–323, December 2003. ISSN 1573-0964. doi: 10.1023/B:SYNT.0000004904.91112.16. URL https://doi.org/10.1023/B:SYNT.0000004904.91112.16.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009.

Judea Pearl. Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes. *Journal of Causal Inference*, 6(2), 2018. doi: 10.1515/jci-2018-2001. URL `https://www.degruyter.com/view/j/jci.2018.6.issue-2/jci-2018-2001/jci-2018-2001.xml`.

Jonas Peters, Dominik Janzing, and Bernard Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.

Peter Selinger. A survey of graphical languages for monoidal categories. *arXiv:0908.3347 [math]*, 813:289–355, 2010. doi: 10.1007/978-3-642-12821-9_4. URL `http://arxiv.org/abs/0908.3347`. arXiv: 0908.3347.

James Woodward. Causation and Manipulability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016. URL `https://plato.stanford.edu/archives/win2016/entries/causation-mani/`.

Karren Yang, Abigail Katoff, and Caroline Uhler. Characterizing and Learning Equivalence Classes of Causal DAGs under Interventions. In *International Conference on Machine Learning*, pages 5537–5546, July 2018. URL `http://proceedings.mlr.press/v80/yang18a.html`.

Appendix: