Causal Statistical Decision Theory|What are interventions?

David Johnston

November 27, 2020

Contents

1 Theories of causal inference

Feedback start here

Beginning in the 1930s, a number of associations between cigarette smoking and lung cancer were established: on a population level, lung cancer rates rose rapidly alongside the prevalence of cigarette smoking. Lung cancer patients were far more likely to have a smoking history than demographically similar individuals without cancer and smokers were around 40 times as likely as demographically similar non-smokers to go on to develop lung cancer. In laborotory experiments, cells which were introduced to tobacco smoke developed *ciliastasis*, and mice exposed to cigarette smoke tars developed tumors(Proctor, 2012). Nevertheless, until the late 1950s, substantial controversy persisted over the question of whether the available data was sufficient to establish that smoking cigarettes *caused* lung cancer. Cigarette manufacturers famously argued against any possible connection (Oreskes and Conway, 2011) and Roland Fisher in particular argued that the available data was not enough to establish that smoking actually caused lung cancer (Fisher, 1958). Today, it is widely accepted that cigarettes do cause lung cancer, along with other serious conditions such as vascular disease and chronic respiratory disease (World Health Organisation, 2018; Wiblin, 2016).

The question of a causal link between smoking and cancer is a very important one. Individuals who enjoy smoking (or think they might) may wish to avoid smoking if cigarettes pose a severe health risk, so they are interested in knowing whether or not it is so. Potential investors in cigarette manufacturers want to know if the product they are backing is likely to see limited adoption due to health concerns. People holding investments in cigarette manufacturering firms want the world to be such that cigarettes do not pose a substantial health risk, as this increases the value of their investment. Governments and organisations with a responsibility for public health may see themselves as having responsibility to discourage smoking as much as possible if smoking is severely detrimental to health. The costs and benefits of poor decisions about smoking are large: 8 million annual deaths are attributed to cigarette-caused cancer and vascular disease in 2018(World Health Organisation, 2018) while global cigarette sales were estimated at US$711 million in 2020, while (Statista, 2020) (a figure which might be substantially larger if cigarettes were not widely believed to be harmful).

The question of whether or not cigarette smoking causes cancer illustrates two key facts about causal questions: First, having the right answers to some causal questions is of tremendous importance to huge numbers of people. Second, even when large amounts of data show unambiguous associations between phenomena of interest, it is still difficult to know when a causal conclusion is justified.

Causal conclusions are often justified on the basis of ad-hoc reasoning. For example Krittanawong et al. (2020) states:

> [...] the potential benefit of increased chocolate consumption, reducing coronary artery disease (CAD) risk is not known. We aimed to explore the association between chocolate consumption and CAD.

It is not clear whether Krittanawong et. al. mean that a negative association between chocolate consumption and CAD implies that increased chocolate consumption is likely to reduce coronary artery disease, or that an association may be relevant to the question and the reader should draw their own conclusions. Whether the implication is being suggested by Krittanawong et. al. or merely imputed by naïve readers, it is being drawn on an ad-hoc basis – no argument for the implication can be found in this paper. As Pearl (2009) has forcefully argued, additional assumptions are always required to answer causal questions from associational facts, and stating these assumptions explicitly allows those assumptions to be productively scrutinised.

Theories of causal inference exist to enable formal rather than ad-hoc reasoning about causal questions. Instead of posing informal causal question and answering them based on ad-hoc reasoning, within a theory of causal inference we ask about properties of "causal models" (which are simply mathematical types defined by the theory) subject to certain assumptions we are willing to make. A successful theory of causal inference should enable causal models that "adequately represent" the original informal question, and the assumptions we invoke should be more accessible to scrutiny than ad-hoc assertions made in the course of answering the informal question.

As well as defining causal models, which represent *claims about causation*, theories of causal inference also formalise the problem of *inferring the correct causal model* - this is the problem of taking some observational data and concluding which causal models are "possible" or "appropriate to use for the given purpose".

Defining causal models is difficult. While philosophical theories of causation are not heavily discussed in the applied literature on causal inference, the principles used to motivate the definitions of causal models are widely discussed in the philosophical literature. Applied theories of causal inference:

1. "X_i causes X_j" means that there exist different *ideal interventions* that result in different values of of X_i, hold other "causally sufficient" variables constant, do not directly affect X_j but nonetheless entail different values of X_j
2. "X_i causes X_j" means that the *counterfactual value* of X_j would be different "if X_i had taken a different value"

In practice, most theories of causal inference seem to be based on the notion of *ideal interventions*. Even "counterfactual" theories of causal inference (such as the theory based on "potential outcomes" notation) tend to define counterfactual values as "values that a variable would have taken were it exposed to an ideal intervention", if they are defined at all (Morgan and Winship, 2014; Rubin, 2005; Richardson and Robins, 2013). There are, however, alternative definitions of counterfactual values such as Lewis' closest world semantics (Lewis, 1986), though these definitions have serious difficulties.

"Ideal interventions" are difficult to define. The structural model approach of Pearl (2009) defines ideal interventions in terms of "causally sufficient models". However, this definition ends up circular:

- An $[\mathsf{X}_i, \mathsf{X}_j]$-ideal intervention is an operation whose result is determined by applying the do-calculus to a causally sufficient triple $((\Omega, \mathcal{F}, \mathbb{Q}), \mathcal{G}, \boldsymbol{U})$
- A triple $((\Omega, \mathcal{F}, \mathbb{Q}), \mathcal{G}, \boldsymbol{U})$ is $[\mathsf{X}_i, \mathsf{X}_j]$-causally sufficient if U contains X_i, X_j and "all intervenable variables" that *cause* (definition (1)) both X_i and X_j [1]

Circularity is a recognised problem with interventional definitions of causation (Woodward, 2016). An alternative approach is to designate certain real-world events – such as flipping coins, querying random number generators and so forth – as prototypical "ideal interventions". The main problem with this definition is that it fails to answer any causal query that hasn't been subject to a randomised trial, which seems to leave most causal questions unanswerable, despite the fact that many of them have apparent answers (Pearl, 2018a), and the causal questions represented under this definition are often considered to be inadequate representations of the original questions to which answers were sought (Deaton and Cartwright, 2018; Heckman, 1991). Additional challenges posed by "ideal interventions" are discussed in Section 4.

An account of ideal interventions has not yet been offered that stands up to scrutiny. This difficulty is not without precedent, and is certainly no reason to abandon theories built upon ideal interventions. It is also difficult to provide an account of what it means for data to be "distributed according to probability distribution $\mathbb{P}$"(Hájek, 2019) that stands up under scrutiny, but the usefulness of probability distributions in representing data regularities is widely accepted.

However, in light of the difficulties with theories of causation, it is notable that many causal questions can be formalised without appealing to a theory

[1] Weaker conditions for causal sufficiency are possible, but they are still premised on causal relationships, so the charge of circularity stands (Shpitser and Pearl, 2008).

of causation at all, an observation also made by Dawid (2020). Causal statistical decision theory (CSDT), introduced here, shows how the type definition of "causal models" is almost completely determined by the given *causal decision problem*. Causal decision problems are problems such as: Should I smoke? Should cigarettes be taxed? Such problems are ubiquitous, and causal inferences are instrumental in answering them. As with existing theories of causal inference, determining the relation between observational probabilities and causal models is the key problem in CSDT, and existing theories of causal inference have developed many useful approaches to this. CSDT additionally offers the opportunity to understand what remains true of causal decision problems regardless of the theory of causation in use and the opportunity to consider theories of causation which may be useful in solving particular causal decision problems but need not be universally valid.

2 Causal Questions

Pearl and Mackenzie (2018) has proposed three types of causal question:

1. Association: How are X and Y related? How would observing X change my beliefs about Y?

2. Intervention: What would happen if I do ... ? How can I make E happen?

3. Counterfactual: Was X responsible for Y? What if I had done ... instead of what I actually did?

I focus on a particular type of question here: "How can I make E happen?". Pearl calls this type of question an "interventional" question but we observe that this query is in fact a *causal decision problem* (I also wish to avoid the term "intervention" as it has a technical meaning which that I touched on above). A causal decision problem is a problem of the following form:

> Given my available options D and the way I would like the world to turn out, which options are likely to have a desirable result?

When one asks "How can I make E happen?", I suppose that the answer will be one or more elements of a set of available options D. Observing that E corresponds to "the way I would like the world to turn out", we can see that this question is a causal decision problem.

I choose to focus on this type of problem because the problem of choosing a decision given desired results is ubiquitous, and Definition 2 is a broadly applicable prototype for this kind of problem. It's possible that a similar approach can succeed for counterfactual queries, but obviously a different prototype problem would be required. Enough questions are raised by focusing on causal decision problems to restrict our attention to this type of problem for now.

Causal *statistical* causal decision problems extend causal decision problems by introducing data and assumptions about how the world works:

> Given my available options D, data $\boldsymbol{\mathsf{X}}$ and preferences u, which options are likely to have a desirable result?

2.1 Formalising Causal Statistical Decision Problems: See-Do Maps

I introduce *see-do maps* which formalise causal statistical decision problems introduced in Definition 2. I don't derive see-do maps from first principles - rather, I make use of probability theory and expected utility theory because they're the standard tools used to formalise models of noisy observations and choice under uncertainty.

Section 8 (currently unfinished) considers more fundamental assumptions that might lead to see-do maps

First, we formally define *causal statistical decision problems.*

Definition 2.1 (Causal Statistical Decision Problem)**.** A *causal statistical decision problem* (CSDP) is a tuple (D, x, X, Y, u) where D is a set of *decisions* that may be taken, $x \in X$ is a *data sequence* of observations, Y is the space of possible *consequences* of decisions and $u : Y \to \mathbb{R}$ is a *utility function* where $u(y) \geq u(y')$ implies that y is at least as desirable as y'.

We place a number of additional requirements on causal statistical decision problems:

- D is a denumerable set
- The principle of *expected utility maximisation* is acceptable, so the desirability of a probability distribution $\mathbb{P}_{\mathsf{Y}} \in \Delta(\mathcal{Y})$ can be found by evaluating $\mathbb{E}_{\mathbb{P}_{\mathsf{Y}}}[u]$
- X and Y are *standard measurable* (D is automatically standard measurable due to it being denumerable)
- A decision maker may select any probability distribution in $\Delta(\mathcal{D})$ as its stochastic decision. A deterministic distribution δ_d is equivalent to directly choosing $d \in D$, while deciding on a non-deterministic distribution involves consulting a source of randomness to select a particular decision in accordance with the probabilities chosen

Definition 2.1 are the elements of a CSDP that we regard as fixed.

A *decision maker* brings prior knowledge to the problem in the form of a *see-do map*, which captures the relation between the *observed probability distribution* and the *consequence map.*

Definition 2.2 (Observed probability distribution)**.** Given a CSDP (D, x, X, Y, u), under hypothesis θ the data x are "distributed according to" some *observed probability* $\mathbb{O}_\theta \in \Delta(\mathcal{X})$.

"Distributed according to" is a weasel phrase with no particularly satisfactory interpretation, but I'd like to avoid worrying about it at this

Definition 2.3 (Consequence map)**.** Given a decision set D and a consequence set Y, under hypothesis θ a *consequence map* $\mathbb{C}_\theta : D \to \Delta(\mathcal{Y})$ is a Markov kernel mapping decisions to probability distributions on the consequence space. For a stochastic decision $\gamma \in \Delta(\mathcal{D})$, $\gamma\mathbb{C}$ is the *consequence* of γ with respect to $\mathbb{C}_\theta$.

Definition 2.4 (Causal hypothesis)**.** A causal hypothesis θ on a CSDP (D, x, X, Y, u) is a pair $(\mathbb{O}_\theta, \mathbb{C}_\theta) \in \Delta(\mathcal{X}) \times \Delta(\mathcal{Y})^D$. A hypothesis $(\mathbb{O}_\theta, \mathbb{C}_\theta)$ can be interpreted as the statement "the observed data are distributed according to $\mathbb{O}_\theta$ and the consequences of decisions are given by $\mathbb{C}_\theta$"

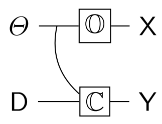
Definition 2.5 (Causal hypothesis class)**.** A *causal hypothesis class* Θ is a set of causal hypotheses on (D, x, X, Y, u).

Definition 2.6 (See-do map)**.** Given a CSDP (D, x, X, Y, u) and a causal hypothesis class Θ, the*see-do* model $\mathbb{T} : \Theta \times D \times (\mathcal{X} \otimes \mathcal{Y})$ is the map $\mathbb{T} : (\theta, d, A \times B) \mapsto \mathbb{O}_\theta(A)\mathbb{C}_{(\theta,d)}(B)$.

We assume that see-do maps are Markov kernels (i.e. there is a σ-algebra $\theta \otimes \mathcal{D}$ such that the map $(\theta, d) \mapsto \mathbb{O}_\theta(A)\mathbb{C}_{(\theta,d)}(B)$ is $\theta \otimes \mathcal{D}$-measurable). This is guaranteed if Θ is denumerable and endowed with the discrete σ-algebra. If Θ is standard measurable, this imposes restrictions on Θ; for example, we can't have some$d \in D$, $A \times B \in \mathcal{X} \otimes \mathcal{Y}$ such that $\mathbb{O}_\cdot(A)\mathbb{C}_{(\cdot,d)}(B)^{-1}(\{1\})$ is a non-measurable collection of hypotheses in Θ.

There is a generic graphical representation of see-do maps. See Section 7.4 for a thorough explanation of the notation used.

Define $\mathbb{O} : \Theta \to \Delta(\mathcal{X})$ by $\mathbb{O} : \theta \mapsto \mathbb{O}_\theta$ and $\mathbb{C} : \Theta \times D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ by $\mathbb{C} : (\theta, d) \mapsto \mathbb{C}_{(\theta,d)}$. Then we have

$\mathbb{T} =$ 

(1)

A key difference between CSDT and other approaches to causal inference is that diagrams in CSDT feature two coupled maps $\mathbb{O}$ and $\mathbb{C}$, while most other approaches to causal inference represent both $\mathbb{O}$ and $\mathbb{C}$ in one diagram. Lattimore and Rohde (2019) is the only other example I am aware of that represents both $\mathbb{O}$ and $\mathbb{C}$. Nevertheless, "one-picture" causal models such as Causal Bayesian Networks, Single World Intervention Graphs *do* represent observational distributions and interventional maps, and the two differ (see Section **??**)

A causal hypothesis class Θ induces a binary relation between observed probability distributions $\mathbb{O}_\theta$ and consequence maps $\mathbb{C}_\theta$. This approach is very agnostic about the actual relation induced – we do not even insist that the range of the observed data X is the same as the range of possible consequences Y (though we will generally limit our attention to cases where the two coincide).

In common with Heckerman and Shachter (1995), decisions (or "acts") are primitive elements of see-do maps. In contrast to our work, Heckerman and

Shachter (1995) only discuss deterministic *consequence maps*, while see-do maps represent relations between consequence maps and observed probability.

Decisions are similar to the "regime indicators" found in Dawid (2020). They coincide precisely if we suppose that the observation and consequence spaces coincide ($X = Y$) and there exists an "idle" decision $d^* \in D$ such that $\mathbb{C}_{(\cdot,d^*)} = \mathbb{O}_{\cdot}$. However, in general we don't require that $\mathbb{O}$ and $\mathbb{C}$ are related in this manner. This assumption will be revisited in

A section I haven't written yet

.

2.2 D-causation

While we take D to be a primitive element of causal decision problems, and therefore a primitive of see-do maps. Causes are not primitive, but we can offer a secondary notion of causation. We call this D-causation to stress the fact that it arises in a theory of causal inference in which the set D of available decisions is primitive. A similar idea is discussed extensively in Heckerman and Shachter (1995). The main differences are that what we call "consequence maps" map decisions to probability distributions over possible consequences while Heckerman and Shachter work with "states" that map decisions deterministically to consequences. In addition, while we define D-causation relative to a particular consequence map $\mathbb{C}_\theta$, Heckerman and Shachter define it with respect to a *set* of states.

Section 4 explores the difficulty of defining "objective causation" without reference to a set of basic decisions, acts or operations. D need not be interpreted as the set of decisions an agent may make, but whatever interpretation it is assigned, all existing examples of causal models seem to require a "domain set".

See Section 7.5 for the definition of random variables.

Add definition of conditional independence, revise wire label definitions

One way to motivate the notion of D-causation is to observe that for many decision problems, the full set D may be extremely large. Suppose I aim to have my light switched on, and there is a switch that controls the light. Often, the relevant choice of acts for such a problem would appear to be $D_0 = \{\text{flip the switch, don't flip the switch}\}$. However, in principle I have a much larger range of options to choose from. For simplicity's sake, suppose I have instead the following set of options:

$$\begin{aligned} D_1 := \{&\text{“walk to the switch and press it with my thumb}'', \\ &\text{“trip over the lego on the floor, hop to the light switch and stab my finger at it}'', \\ &\text{“stay in bed}''\} \end{aligned}$$

If having the light turned on is all that matters, I could consider any acts in D_1 to be equivalent if they have the same ultimate impact on the position of the light switch. D_0 is a quotient over D_1 under this equivalence relation.

If I hypothesize that, relative to D_1, the ultimate state of the light switch is all that matters to determine the ultimate state of the light, I can say that the light switch D_1-causes the state of the light. Given this D_1-causation, the D_1 decision problem can (subject to my hypothesis) be reduced to a D_0 decision between states of the light switch.

If I consider an even larger set of possible acts D_2, I might not accept the hypothesis of D_2-causation. Let D_2 be the following acts:

D_2 :={“walk to the switch and press it with my thumb″,
“trip over the lego on the floor, hop to the light switch and stab my finger at it″,
“stay in bed″, “toggle the mains po

In this case, it would be unreasonable to hypothesize that all acts that left the light switch in the “on” position would also result in the light being “on”. Thus the switch does not D_2-cause the light to be on.

Formally, D-causation is defined in terms of conditional independence:

Definition 2.7 (D-causation)**.** Given a consequence map $\mathbb{C}_\theta : D \to \Delta(\mathcal{Y})$, random variables $\mathsf{Y}_1 : Y \times D \to Y_1$, $\mathsf{Y}_2 : Y \times D \to Y_2$ and domain variable $\mathsf{D} : Y \times D \to D$ (Definition 7.8), Y_1 D-causes Y_2 iff $\mathsf{Y}_2 \perp\!\!\!\perp_{\mathbb{C}_\theta} \mathsf{D}|\mathsf{Y}_1$.

2.3 D-causation vs Heckerman and Shachter

Heckerman and Shachter study deterministic “consequence maps”. Furthermore, what we call hypotheses $\theta \in \Theta$, Heckerman and Schachter call states $s \in S$. One could consider a state to be a hypothesis that is specific enough to yield a deterministic map from decisions to outcomes. Heckerman and Shachter’s notion of causation is defined by *limited unresponsiveness* rather than *conditional independence*, which depends on a partition of states rather than a particular hypothesis.

Definition 2.8 (Limited unresponsiveness)**.** Given states S, deterministic consequence maps $\mathbb{C}_s : D \to \Delta(F)$ for each $s \in A$ and a random variables $\mathsf{X} : F \to X$, $\mathsf{Y} : F \to Y$, Y is unresponsive to D in states limited by X if $\mathbb{C}^{\mathsf{X}|\mathsf{D}}_{(s,d)} = \mathbb{C}^{\mathsf{X}|\mathsf{DS}}_{(s,d')} \implies \mathbb{C}^{\mathsf{Y}|\mathsf{DS}}_{(s,d)} = \mathbb{C}^{\mathsf{Y}|\mathsf{DS}}_{(s,d')}$ for all $d, d' \in D$, $s \in S$. Write $\mathsf{Y} \not\hookleftarrow_{\mathsf{X}} \mathsf{D}$

Lemma 2.9 (Limited unresponsiveness implies D-causation)**.** *For deterministic consequence maps,* $\mathsf{Y} \not\hookleftarrow_{\mathsf{X}} \mathsf{D}$ *implies* X *D-causes* Y *in every state $s \in S$.*

Proof. By the assumption of determinism, for each $s \in S$ and $d \in D$ there exists $x(s, d)$ and $y(s, d)$ such that $\mathbb{C}^{\mathsf{XY}|\mathsf{DS}}_{d,s} = \delta_{x(s,d)} \otimes \delta_{y(s,d)}$.

By the assumption of limited unresponsiveness, for all d, d' such that $x(s, d) = x(s, d')$, $y(s, d) = y(s, d')$ also. Define $f : X \times S \to Y$ by $(s, x) \mapsto y(s, [x(s, \cdot)]^{-1}(x(s, d)))$ where $[x(s, \cdot)]^{-1}(a)$ is an arbitrary element of $\{d|x(s, d) = a\}$. For all s, d, $f(x(s, d), s) = y(s, d)$. Define $\mathbb{M} : X \times D \times S \to \Delta(\mathcal{Y})$ by $(x, d, s) \mapsto \delta_{f(x,s)}$. $\mathbb{M}$ is a version of $\mathbb{C}^{\mathsf{Y}|\mathsf{X},\mathsf{D},\mathsf{S}}$ because, for all $A \in \mathcal{X}$, $B \in \mathcal{Y}$, $s \in S$, $d \in D$:

$$\mathbb{C}^{\mathsf{X}|\mathsf{DS}}_{(d,s)}\curlyvee(\mathbb{M}\otimes\mathrm{Id}) = \int_A \mathbb{M}(x',d,s;B)d\delta_{x(s,d)}(x') \tag{2}$$

$$= \int_A \delta_{f(x',s)}(B)d\delta_{x(s,d)}(x') \tag{3}$$

$$= \delta_{f(x(s,d),s)}(B)\delta_{x(s,d)}(A) \tag{4}$$

$$= \delta_{y(s,d)}(B)\delta_{x(s,d)}(A) \tag{5}$$

$$= \delta_{x(s,d)}\otimes\delta_{y(s,d)}(A\times B) \tag{6}$$

$\mathbb{M}$ is also independent of D, given the obvious labeling of inputs. Therefore $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{C}_s} \mathsf{D}|\mathsf{X}$. □

However, despite limited unresponsiveness implying D-causation within every state, it does not imply D-causation in mixtures of states. Suppose $D = \{0,1\}$ where 1 stands for "toggle light switch" and 0 stands for "do nothing". Suppose $S = \{[0,0],[0,1],[1,0],[1,1]\}$ where $[0,0]$ represents "switch initially off, mains off" the other states generalise this in the obvious way. Finally, $\mathsf{F}\in\{0,1\}$ is the final position of the switch and $\mathsf{L}\in\{0,1\}$ is the final state of the light. We have

$$\mathbb{C}^{\mathsf{LF}|\mathsf{DS}}_{d,[i,m]} = \delta_{(d \text{ XOR } i) \text{ AND } m}\otimes\delta_{(d \text{ XOR } i) \text{ AND } m} \tag{7}$$

Within states $[0,0]$ and $[1,0]$, the light is always off, so $\mathsf{F} = a \implies \mathsf{L} = 0$ for any a. In states $[0,1]$ and $[1,1]$, $\mathsf{F} = 1 \implies \mathsf{L} = 1$ and $\mathsf{F} = 0 \implies \mathsf{L} = 0$. Thus $\mathsf{L} \not\hookleftarrow_{\mathsf{F}} \mathsf{D}$. However, suppose we take a mixture of consequence maps:

$$\mathbb{C}_\gamma = \frac{1}{4}\mathbb{C}_{\cdot,[0,0]} + \frac{1}{4}\mathbb{C}_{\cdot,[0,1]} + \frac{1}{2}\mathbb{C}_{\cdot,[1,1]} \tag{8}$$

$$\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma = \frac{1}{4}\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\otimes\begin{bmatrix}1 & 0\\1 & 0\end{bmatrix} + \frac{1}{4}\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\otimes\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}0 & 1\\1 & 0\end{bmatrix}\otimes\begin{bmatrix}0 & 1\\1 & 0\end{bmatrix} \tag{9}$$

Then

$$[1,0]\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma = \frac{1}{4}[0,1]\otimes[1,0] + \frac{1}{4}[0,1]\otimes[0,1] + \frac{1}{2}[1,0]\otimes[1,0] \tag{10}$$

$$[1,0]\curlyvee(\mathbb{C}^{\mathsf{F}|\mathsf{D}}_\gamma\otimes\mathbb{C}^{\mathsf{L}|\mathsf{D}}_\gamma) = (\frac{1}{2}[0,1]+\frac{1}{2}[1,0])\otimes(\frac{1}{4}[0,1]+\frac{3}{4}[1,0]) \tag{11}$$

$$\implies [1,0]\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma \neq [1,0]\curlyvee(\mathbb{C}^{\mathsf{F}|\mathsf{D}}_\gamma\otimes\mathbb{C}^{\mathsf{L}|\mathsf{D}}_\gamma) \tag{12}$$

Thus under hypothesis mixture γ, F does not D-cause L even though F D-causes L in all states S. The definition of D-causation was motivated by the idea that we could reduce a difficult decision problem with a large set D to a simpler problem with a smaller "effective" set of decisions by exploiting conditional independence. Even if X D-causes Y in every $\theta\in S$, X does not necessarily D-cause Y in mixtures of states in S. For this reason, we do not say that X D-causes Y in S if X D-causes Y in every $\theta \in S$, and in this way we differ substantially from Heckerman and Shachter (1995).

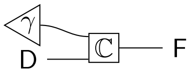
Instead, we simply extend the definition of D-causation to mixtures of hypotheses: if $\gamma \in \mathit{\Delta}(\mathit{\Theta})$ is a mixture of hypotheses, define $\mathbb{C}_\gamma := (\gamma \otimes \mathbf{Id})\mathbb{C}$. Then X D-causes Y relative to γ iff $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{C}_\gamma} \mathsf{D}|\mathsf{X}$.

Theorem 2.10 shows that under some conditions, D-causation can hold for arbitrary mixtures over subsets of the hypothesis class $\mathit{\Theta}$.

Theorem 2.10 (Universal D-causation)**.** *If* $\mathbb{C}_\theta^{\mathsf{X}|\mathsf{D}} = \mathbb{C}_{\theta'}^{\mathsf{X}|\mathsf{D}}$ *for all* $\theta, \theta' \in S \subset \mathit{\Theta}$ *and* X *D-causes* Y *in all* $\theta \in S$*, then* X *D-causes* Y *with respect to all mixed consequence maps* $\mathbb{C}_\gamma$ *for all* $\gamma \in \mathit{\Delta}(\mathit{\Theta})$ *with* $\gamma(S) = 1$*.*

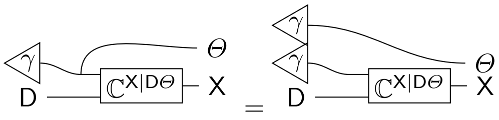
Proof. For $\gamma \in \mathit{\Delta}(\mathit{\Theta})$, define the mixture

$\mathbb{C}_\gamma :=$



(13)

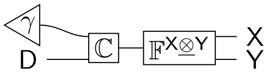
Because $\mathbb{C}_\theta^{\mathsf{X}|\mathsf{D}} = \mathbb{C}_{\theta'}^{\mathsf{X}|\mathsf{D}}$ for all $\theta, \theta' \in \mathit{\Theta}$, we have


γ
$\mathit{\Theta}$
D
$\mathbb{C}^{\mathsf{X}|\mathsf{D}\mathit{\Theta}}$
X
=
γ
γ
$\mathit{\Theta}$
D
$\mathbb{C}^{\mathsf{X}|\mathsf{D}\mathit{\Theta}}$
X

(14)

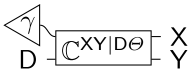
Also

$\mathbb{C}_\gamma^{\mathsf{XY|D}} =$



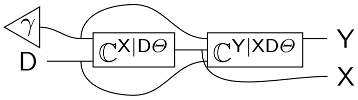
(15)

$=$



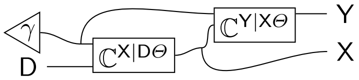
(16)

$=$



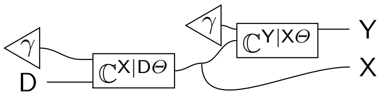
(17)

$\overset{\mathsf{Y}\perp\!\!\!\perp\mathsf{D}|\mathsf{X}\Theta}{=}$



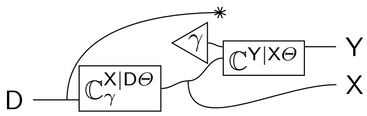
(18)

$\overset{14}{=}$



(19)

$=\overset{14}{=}$



(20)

Equation 20 establishes that $(\gamma \otimes \mathbf{Id}_X \otimes *_D)\mathbb{C}^{\mathsf{Y|X}\Theta}$ is a version of $\mathbb{C}_\gamma^{\mathsf{Y|XD}}$, and thus $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{C}_\gamma} \mathsf{D}|\mathsf{X}$.

This can also be derived from the semi-graphoid rules:

$$\Theta \perp\!\!\!\perp \mathsf{D} \wedge \Theta \perp\!\!\!\perp \mathsf{X}|\mathsf{D} \implies \Theta \perp\!\!\!\perp \mathsf{XD} \tag{21}$$

$$\implies \Theta \perp\!\!\!\perp \mathsf{D}|\mathsf{X} \tag{22}$$

$$\mathsf{D} \perp\!\!\!\perp \Theta|\mathsf{X} \wedge \mathsf{D} \perp\!\!\!\perp \mathsf{Y}|\mathsf{X}\Theta \implies \mathsf{D} \perp\!\!\!\perp \mathsf{Y}|\mathsf{X} \tag{23}$$

$$\implies \mathsf{Y} \perp\!\!\!\perp \mathsf{D}|\mathsf{X} \tag{24}$$

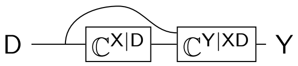
□

2.4 Properties of D-causation

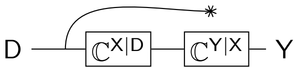
If X D-causes Y relative to $\mathbb{C}_\theta$, then the following holds:

$$\mathbb{C}_\theta^{\mathsf{X|D}} = \mathsf{D} - \boxed{\mathbb{C}^{\mathsf{X|D}}} - \boxed{\mathbb{C}^{\mathsf{Y|X}}} - \mathsf{Y} \tag{25}$$

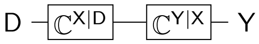
This follows from version (2) of Definition 7.28:

$\mathbb{C}_\theta^{\mathsf{X}|\mathsf{D}} =$  (26)

D
$\mathbb{C}^{\mathsf{X}|\mathsf{D}}$
$\mathbb{C}^{\mathsf{Y}|\mathsf{XD}}$
Y

$=$  (27)

D
$\mathbb{C}^{\mathsf{X}|\mathsf{D}}$
$\mathbb{C}^{\mathsf{Y}|\mathsf{X}}$
Y

$=$  (28)

D
$\mathbb{C}^{\mathsf{X}|\mathsf{D}}$
$\mathbb{C}^{\mathsf{Y}|\mathsf{X}}$
Y

D-causation is not transitive: if X D-causes Y and Y D-causes Z then X doesn't necessarily D-cause Z.

3 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 $\mathsf{\mathit{X}}$
- 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 7.9
- Given a random variable X, F_X is the associated Markov kernel, see Definition 7.1
- See section 7.4.2 for rules of string diagram manipulation
- $*_X : X \to \Delta(\{*\})$ is the discard map defined in Equation 147

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

See-Do models and Causal Bayesian Networks (and related models such as SCMs) are quite different in their appearance and in the interpretation of various elements. In defining See-Do models, we assume that there is a decision problem that fixes in advance the observation space E, the space of consequences F and the set of available decisions D. By including D, See-Do models have an agent "baked in" to the definition. In contrast, Causal Bayesian Networks assume that a set of observed variables $\mathbf{X}$ and a set of unobserved variables $\mathbf{U}$ is fixed by nature. A subtlety here is that $\mathbf{U}$ is generally *not* known, but it may be assumed that whatever variables actually comprise $\mathbf{U}$, they may be generically representable by some known set of variables $\mathbf{U}'$ without loss. Because a set of decisions D is absent, Causal Bayesian Networks appear to model agent-independent "causal relationships".

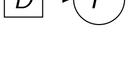
Despite the fact that Causal Bayesian Networks don't seem to built to model the consequences of an agent's decision making, they are nonetheless considered to be appropriate for this purpose. This is because the "do-operations" that Causal Bayesian Networks support are considered to have some relationship to any set D decisions some agent might want to consider.

The question of how this relationship might be determined in general is one that I have not seen addressed anywhere. Typically, the approach taken is "I know it when I see it". For example, if I were a doctor and I could either A) hand a patient a prescription or B) not hand the patient the prescription, and if I had a set of observational data of past patients including a variable S representing whether or not they had received a prescription for the drug, I could consider option A to correspond to $do(\mathsf{S} = 1)$ and option B to correspond to $do(\mathsf{S} = 0)$ in some causal model (perhaps in the "true" causal model). While this might appear to be reasonable, we should be cautious: this is a completely *ad-hoc* assumption.

In fact, given a sufficiently rich set of variables $\mathbf{U}\cup\mathsf{X}$, I argue that $do(\mathsf{S})$ will almost never even approximate the consequences of an action known in advance to fix the value of some variable S. For this reason, it is valuable to have a theory like CSDT that is concerned only with the consequences of actions.

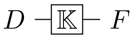
4.1 Influence Diagrams vs See-Do models

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:



(29)

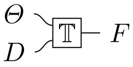
Is an influence diagram with the decision node D and the chance node F. When nodes are associated with sets representing possible values, influence diagrams represent sets of Markov kernels. For example, if we associate the measurable set $(D, \mathcal{D})$ with D and $(F, \mathcal{F})$ with F, then we could take Diagram 29 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:



(30)

This diagram represents a *particular* Markov kernel $\mathbb{K} : D \to \Delta(\mathcal{F})$. A set such as that represented by Diagram 29 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 29 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

notation



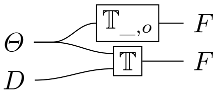
(31)

Diagram 31 has a few more elements that Diagram 29 - Θ and $\mathbb{T}$ in particular. If I were to define exactly what D, Θ, $\mathbb{T}$ and F were, Diagram 31 would represent a unique Markov kernel. On the other hand, Diagram 29 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 29 along with additional assumptions that represents "the same thing".

Under the condition that the choice of index "doesn't matter" somehow

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

notation



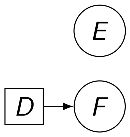
(32)

While this is a feature of influence diagrams in Dawid (2002), *in general* a diagram like 29 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.

There is some kind of measurability condition needed to make the two representations coincide

Concretely, given the influence diagram

$$\mathcal{J} =$$

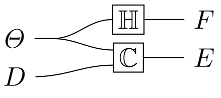


(33)

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 $(\mathcal{J}, A)$ 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

$$\mathbb{T} :=$$



(34)

Maybe move the following to an appendix?

Definition 4.1 (Markov kernel/influence diagram compatibility)**.** Given a Markov kernel $\mathbb{K} : E \to \Delta(\mathcal{F})$, an influence diagram $\mathcal{J} = (S, A, E)$ 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 $X_1 \perp_{\mathcal{J}} X_2 \implies f(\mathsf{X}_1) \perp\!\!\!\perp_K f(\mathsf{X}_2)$

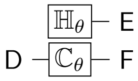
still need to *define* conditional independence WRT kernels, though I think I can do so

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 $\mathcal{J} = (\{E, F\}, \{D\}, \{D \rightarrow F\})$ with respect to the injective function

$$f : \begin{cases} A \mapsto \mathsf{D} \\ E \mapsto \mathsf{E} \\ F \mapsto \mathsf{F} \end{cases} \tag{35}$$

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

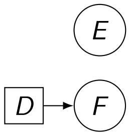
$$\mathbb{T}_\theta =$$



(36)

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

The influence diagram

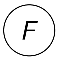
$\mathcal{J} =$ 

(37)

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

4.2 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 29, a Causal Bayesian Network for the same system might be



(38)

The difference between Diagram 29 and Diagram 38 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 38 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.

4.3 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

4.4 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 4.11 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.

4.5 Recursive Structural Causal Models

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

Definition 4.2 (Recursive Structural Causal Model)**.** A recursive structural causal model (SCM) is a tuple

$$\mathcal{M} := \langle N, M, \mathbf{X}_{[N]}, \mathbf{E}_{[M]}, \{f_i | i \in [N]\}, \mathbb{P}_{\mathcal{E}} \rangle \tag{39}$$

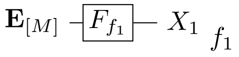
where

- $N \in \mathbb{N}$ 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*

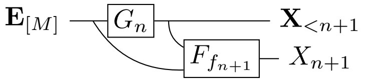
- $\mathbb{P}_{\mathcal{E}} \in \Delta(\mathbf{E}_{\mathcal{J}})$ is a probability measure on the space of exogenous variables

Definition 4.3 (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:

$G_1 =$ (40)



$G_{n+1} =$ (41)



Definition 4.4 (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 7.3). 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 4.5, $\underline{\otimes}_{i \in [N]} \mathsf{X}_i = \mathrm{Id}_{\mathbf{X}_{[N]}}$, 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 4.5 (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 $\mathbf{X} \in \mathbf{X}_{[N]}$,

$$\underline{\otimes}_{i \in [N]} \mathsf{X}_i(\mathbf{X}) = (\pi_1(\mathbf{X}), ..., \pi_N(\mathbf{X})) \tag{42}$$
$$= (x_1, ..., x_n) \tag{43}$$
$$= \mathbf{X} \tag{44}$$

□

Definition 4.6 (Hard Interventions)**.** Let $\mathfrak{M}$ 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 $Do_{\mathcal{S}} : \mathbf{X}_{\mathcal{S}} \times \mathfrak{M} \to \mathfrak{M}$ such that for $\mathbf{a} \in \mathbf{X}_S$, $Do_{\mathcal{S}}(\mathbf{a}, \mathcal{M}) = \langle N, M, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{[M]}, \{f'_i | i \in [N]\}, \mathbb{P}_{\mathcal{E}} \rangle$ where

$$f'_i = f_i \qquad i \notin \mathcal{S} \tag{45}$$
$$f'_i = \pi_i(\mathbf{a}) \qquad i \in \mathcal{S} \tag{46}$$

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

4.6 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

$$\mathsf{X}_3 = \frac{\mathsf{X}_1}{\mathsf{X}_2} \tag{47}$$

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

Definition 4.7 (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*.

We can recover regular SCMs by imposing only trivial constraints:

Lemma 4.8 (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 4.7, 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. □

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

$$\begin{align} \mathsf{X}_{O\setminus\{j\}} &= \mathbf{b} \tag{48}\\ \implies \mathsf{X}_j &= a \tag{49}\end{align}$$

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:

$$\begin{align} [\mathsf{X}_1, \mathsf{X}_2] &= [b_1, b_2] \tag{50}\\ \implies \mathsf{X}_3 &= \frac{b_1}{b_2} \tag{51}\\ [\mathsf{X}_2, \mathsf{X}_3] &= [b_2, b_3] \tag{52}\\ \implies \mathsf{X}_1 &= b_2 b_3 \tag{53}\\ [\mathsf{X}_1, \mathsf{X}_3] &= [b_1, b_3] \tag{54}\\ \implies \mathsf{X}_2 &= \frac{b_1}{b_3} \tag{55}\end{align}$$

The following is a generally useful lemma that should probably be in basic definitions of Markov kernel spaces

Lemma 4.9 (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:

$$S_i^O = \begin{cases} \mathrm{Id}_{X_i} & i \in O \\ \ast_{X_i} & i \notin O \end{cases} \tag{56}$$

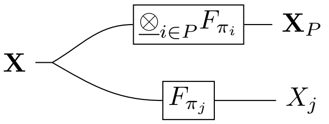
Then $\underline{\otimes}_{i\in O} F_{\pi_i} = \otimes_{i\in\mathcal{I}} S_i^O$.

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 $\ast_{\mathbf{X}}$.

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

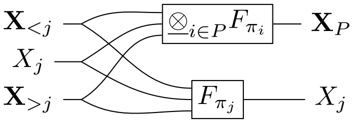
Suppose for $P \subsetneq O$ with greatest element k we have $\underline{\otimes}_{i\in P} F_{\pi_i} = \otimes_{i\in\mathcal{I}} S_i^P$, and suppose that j is the next element of O not in P.

$(\underline{\otimes}_{i\in P} F_{\pi_i}) \underline{\otimes} F_{\pi_j} =$



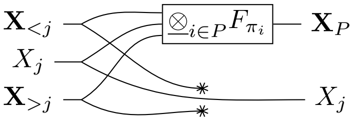
(57)

$=$



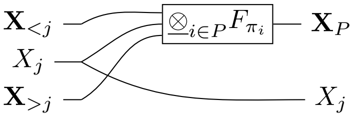
(58)

$=$



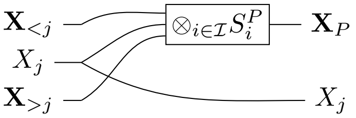
(59)

$=$



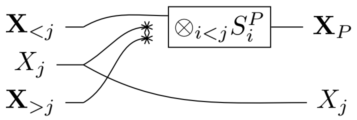
(60)

$=$

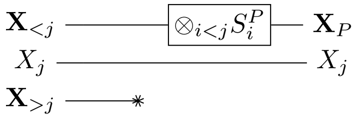


(61)

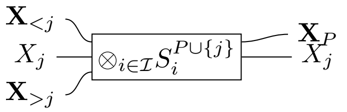
Because all elements of P are less than j, the selector S_k^P resolves to the discard map for $k > j$:



$=$ (62)



$=$ (63)



$=$ (64)

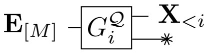
Where 64 follows from the definition of the selector $S_i^{P\cup\{j\}}$.

The proof follows by induction on the elements of O.

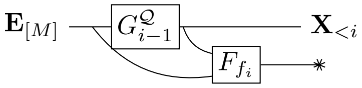
□

Lemma 4.10 (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 40 and 41 with respect to model $\mathcal{Q}$. Note that from Equation 41



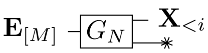
$=$



(65)

$$= G_{i-1}^{\mathcal{Q}} \tag{66}$$

It follows that



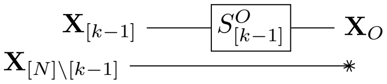
$$= G_{i-1} \tag{67}$$

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

$$G_i^{\mathcal{M}} = G_i^{\mathcal{M}^{do(\mathsf{X}_k=a)}} \tag{68}$$

for all $i < k$. By lemma 4.9, 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 $F_{\mathsf{X}_O} = (\otimes_{i\in[k-1]} S_i^O) \otimes ⁜_{\mathbf{X}_{[N]\setminus[k-1]}}$. Defining $S_{[k-1]}^O := \otimes_{i\in[k-1]} S_i^O$, we have:

$F_{\mathsf{X}_O} =$

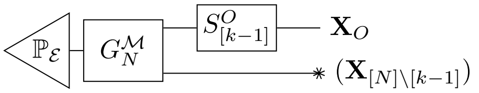


(69)

Thus

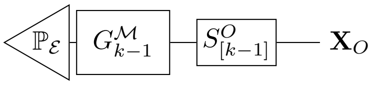
$$\mathbb{P}_{\mathcal{M}}(\mathsf{X}_O) = \mathbb{P}_{\mathcal{E}} G_N^{\mathcal{M}} F_{\mathsf{X}_O} \tag{70}$$

$\overset{69}{=}$


$\mathbb{P}_{\mathcal{E}}$
$G_N^{\mathcal{M}}$
$S^O_{[k-1]}$
$\mathbf{X}_O$
$(\mathbf{X}_{[N]\setminus[k-1]})$

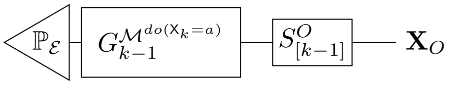
(71)

$\overset{67}{=}$



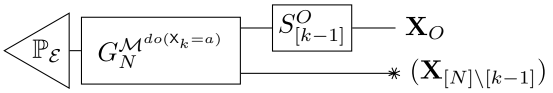
(72)

$\overset{68}{=}$


$\mathbb{P}_{\mathcal{E}}$
$G_{k-1}^{\mathcal{M}^{do(\mathsf{X}_k=a)}}$
$S^O_{[k-1]}$
$\mathbf{X}_O$

(73)

$\overset{67}{=}$


$\mathbb{P}_{\mathcal{E}}$
$G_N^{\mathcal{M}^{do(\mathsf{X}_k=a)}}$
$S^O_{[k-1]}$
$\mathbf{X}_O$
$(\mathbf{X}_{[N]\setminus[k-1]})$

(74)

$$= P_{\mathcal{M}^{do(\mathsf{X}_k=a)}}(\mathsf{X}_O) \tag{75}$$

□

Theorem 4.11 (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 $B^A = \{\mathbf{b} \in \mathbf{X}_{O\setminus k} | \mathsf{X}_{O\setminus k} = \mathbf{b} \implies \mathsf{X}_k \in A\}$ and $B^{A^C} = \{\mathbf{b} \in \mathbf{X}_{O\setminus k} | \mathsf{X}_{O\setminus k} = \mathbf{b} \implies \mathsf{X}_k \notin A\}$.

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

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

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

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

By additivity agian

$$
\begin{aligned}
\mathbb{P}_{\mathcal{M}}\left(\mathsf{X}_{O\setminus\{k\}} \notin B^{A} \ \ \& \ \ \mathsf{X}_k \in A\right) &= \mathbb{P}_{\mathcal{M}}\left(\mathsf{X}_{O\setminus\{k\}} \in B^{A^C} \ \ \& \ \ \mathsf{X}_k \in A\right) && (76)\\
&\quad + \mathbb{P}_{\mathcal{M}}\left(\mathsf{X}_{O\setminus\{k\}} \in (B^{A^C} \cup B^{A})^C \ \ \& \ \ \mathsf{X}_k \in A\right) && (77)\\
&<= 0 + P_{\mathcal{M}}\left(\mathsf{X}_{O\setminus\{k\}} \in (B^{A^C} \cup B^{A})^C\right) && (78)\\
&= 0 && (79)
\end{aligned}
$$

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

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, $\mathbb{P}_{\mathcal{M}^{do(\mathsf{X}_k=a)}}(\mathsf{X}_{O\setminus\{k\}} \in B^{A^C}) = \mathbb{P}_{\mathcal{M}^{do(\mathsf{X}_k=a)}}(\mathsf{X}_k \in A^C)$.

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 4.10, $\mathbb{P}_{\mathcal{M}^{do(\mathsf{X}_k=a)}}(\mathsf{X}_{O\setminus\{k\}} \in B^{A^C}) = P_{\mathcal{M}}(\mathsf{X}_{O\setminus\{k\}} \in B^{A^C}) = P_{\mathcal{M}}(\mathsf{X}_k \in A^C) > 0$. 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. □

4.7 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 4.12 (Structural Causal Model)**.** A structural causal model (SCM) is a tuple

$$\mathcal{M} := \langle \mathcal{I}, \mathcal{J}, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{\mathcal{J}}, \mathbf{f}_{\mathcal{I}}, \mathbb{P}_{\mathcal{E}}, \boldsymbol{E}_{\mathcal{J}} \rangle \tag{80}$$

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
- $\mathbf{f}_{\mathcal{I}} = \underline{\otimes}_{i\in\mathcal{I}} f_i$ 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
- $\boldsymbol{\mathsf{E}}_{\mathcal{J}} = \underline{\otimes}_{j\in\mathcal{J}} \mathsf{E}_j$ is the set of exogenous variables, with $\mathbb{P}_{\mathcal{E}} = \boldsymbol{\mathsf{E}}_{\mathcal{J}\#} P_{\mathcal{E}}$ 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

$$\mathbf{X} = \mathbf{f}_{\mathcal{I}}(\mathbf{X}, \mathbf{e}) \tag{81}$$

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):

$$\mathsf{X}_i = f_i(\boldsymbol{\mathsf{X}}_{\mathcal{I}}, \boldsymbol{\mathsf{E}}_{\mathcal{J}}) \tag{82}$$

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

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

Definition 4.13 (Mechanism surgery)**.** Let $\mathcal{M}$ be the set of SCMs with elements $\langle \mathcal{I}, \mathcal{J}, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{\mathcal{J}}, \_, \mathbb{P}_{\mathcal{E}}, \boldsymbol{\mathsf{E}}_{\mathcal{J}} \rangle$ (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 $\mathcal{M}' = \langle \mathcal{I}, \mathcal{J}, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{\mathcal{J}}, \mathbf{f}'_{\mathcal{I}}, \mathbb{P}_{\mathcal{E}}, \boldsymbol{\mathsf{E}}_{\mathcal{J}} \rangle$.

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 $\mathbb{P}_{\dots}(\dots)$ to specify exactly which random variables we are talking about.

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

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.

4.8 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, 2018b)

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

4.8.1 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 4.14 (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* $\mathbf{g}_\mathcal{I} := \underline{\bigotimes}_{i\in\mathcal{I}} g_i$ 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

$$\mathbf{X} = \mathbf{f}_{\mathcal{I}}(\mathbf{X}, \mathbf{e}) \tag{83}$$

$$\mathbf{X} = \mathbf{g}_{\mathcal{I}}(\mathbf{X}) \tag{84}$$

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 4.15 (Marginal model)**.** Given an SCNM

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

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

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

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

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

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 4.16 (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

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

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

However, as Theroem 4.19 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 4.18 (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

$$\mathcal{I}' = \mathcal{I} \cup \{i'\} \tag{85}$$

$$\mathbf{X}_{\mathcal{I}'} = \mathbf{X}_\mathcal{I} \cup X_{i'} \tag{86}$$

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

$$\mathbf{X}' = \mathbf{f}'_{\mathcal{I}'}(\mathbf{X}', \mathbf{e}) \tag{87}$$

$$\text{and } \mathbf{X}' = \mathbf{g}'_{\mathcal{I}'}(\mathbf{X}') \implies \pi_{\mathcal{I}'\setminus i'}(\mathbf{X}') \qquad = \mathbf{f}_\mathcal{I}(\pi_{\mathcal{I}'\setminus i'}(\mathbf{X}'), \mathbf{e}) \tag{88}$$

$$\text{and } \pi_{\mathcal{I}'\setminus i'}(\mathbf{X}') = \mathbf{g}'_{\mathcal{I}'}(\pi_{\mathcal{I}'\setminus i'}(\mathbf{X}')) \tag{89}$$

Theorem 4.19 (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

$$\mathbf{X}^e = \mathbf{f}_\mathcal{I}(\mathbf{X}^e, \mathbf{e}) \tag{90}$$

$$\mathbf{X}^e = \mathbf{g}_\mathcal{I}(\mathbf{X}^e) \tag{91}$$

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

$$x'^e_{i'} = \mathbf{g}_{i'}(\mathbf{X}'^e) \tag{92}$$

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

$$\mathbf{g}_{i'}(\mathbf{X}'^{e_a}) \neq \mathbf{g}_{i'}(\mathbf{X}'^{e_b}) \tag{93}$$

Therefore there exists no $a \in X_{i'}$ that can simultaneously satisfy 92 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. □

Corollary 4.20. *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 4.19 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 4.19 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

And necessary? There might be "degenerate" necessary relationships that don't harm the possibility of defining interventions, and I'd need to show an equivalence to an SCM in this case

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 4.21 (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 $\mathcal{M}_{\text{atom}}$.

Typically, in order to get an actually usable model you'll also need to marginalize, but I think this complication can be avoided

Definition 4.22 (Causal universality hypothesis)**.** There exists a uniquely solvable SCM $\mathcal{M}_{\text{atom}}$ which is the atomic model that correctly represents all decision problems

what does that mean?

or causal problems?

I don't know how to define "correctly represents" or "causal problem", but it seems like something like the universality hypothesis is necessary if you want to define "the causal effect of X" independent of any atomic model

Relate decisions to interventions on atomic model. Decisions->atomic model is straightforward, but the reverse direction is not so obvious

Causal effects are uniquely defined via atoms iff they are defined via decisions

Are there any plausible ways to construct atomic models?

5 Potential Outcomes

Counterfactual models would be easy to understand if you'd never heard of the Potential Outcomes approach to causality, but learning Potential Outcomes makes them harder to understand. In the following section, I attempt to untangle this mess.

Potential Outcomes is an approach to formalising causal questions. Like Causal Bayeisan Networks, and unlike CSDT, it is a "causes first" approach. Potential Outcomes is motivated by the counterfactual definition of causation - that the "causal effect" of X on Y is the answer to the question "what *would have* happened to Y were X different?". While CSDT was motivated by the question of "what *will* happen to Y if I take some action?", see-do models are perfectly capable of modelling counterfactuals. Instead of D representing *decisions* or *acts*, we can interpret D as representing the available *suppositions* or *counterfactual acts*. Under this shift in interpretation Potential Outcomes models are a generalisation of see-do models.

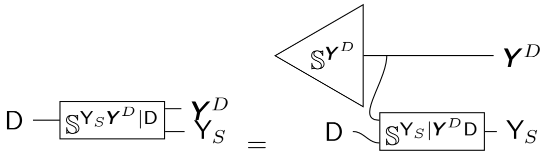
The core of any Potential Outcomes model is the potential outcomes. A standard definition of potential outcomes is: given a random variable Y, potential outcomes of Y with respect to the set of viable suppositions D are random variables Y^d, $d \in D$ which share a codomain with Y and each potential outcome Y^d represents "the value that 'Y' would take *supposing* d". This definition is problematic as the object referred to by 'Y' in the quoted definition is not the random variable Y – were they the same thing, they would always take the same value, which defeats the whole point of supposing things. A good definition of potential outcomes must define 'Y'.

To make sense of it, think about what is meant by "suppose d". We can suppose any $d \in D$ and after supposing d we have some idea of what the world would be like if d were so. That is, supposing is a function from D to "states of the world". We have chosen to represent "states of the world" as probability measures on the outcome space $(E, \mathcal{E})$, so we might as well say supposing is a map $\mathbb{S} : D \to Delta(E)$. The aforementioned '$\mathsf{Y}$' is therefore a random variable on the kernel space $(\mathbb{S}, D \times E)$, which we will call Y_S. Now we have a precise definition of Y_S, but we still need to account for the relation between Y_S and the potential outcomes Y^d stipulated in the original definition.

The relation between Y_S and the potential outcomes requires the assumption of *supposition stability*, which is my own: *the joint distribution of observed variables and potential outcomes is the same under any supposition.* That is, $\underline{\bigotimes}_{d\in D} \mathsf{Y}^d \perp\!\!\!\perp_{\mathbb{C}} \mathsf{D}$, with D being the domain random variable as defined in 7.8.

This assumption looks like "ignorability", but it isn't, and the difference will be explained shortly.

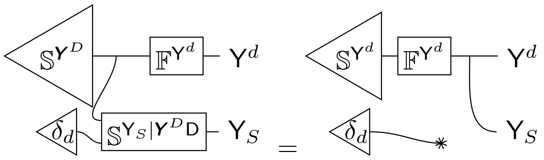
Define $\boldsymbol{Y}^D := \underline{\bigotimes}_{d\in D}\mathsf{Y}^d$. Given the assumption of supposition stability, we can factorise $\mathbb{S}$ as follows:


D
$\mathbb{S}^{\mathsf{Y}_S\boldsymbol{Y}^D|\mathsf{D}}$
$\boldsymbol{Y}^D$
Y_S
=
$\mathbb{S}^{\boldsymbol{Y}^D}$
$\boldsymbol{Y}^D$
D
$\mathbb{S}^{\mathsf{Y}_S|\boldsymbol{Y}^D\mathsf{D}}$
Y_S

(94)

Where $\mathbb{S}^{\mathsf{Y}_S|\mathsf{D}} = \mathbb{S}\mathbb{F}^{\mathsf{Y}_S}$ is the marginal of Y_S and $\mathbb{S}^{\boldsymbol{Y}^D}$ is the marginal of $\boldsymbol{Y}^D$, a version of which (independe of D) exists by supposition stability.

The original definition of a potential outcome says that the value of Y^d is the same as the value of Y_C under supposition d. This is equivalent to requiring for all $d \in D$:


$\mathbb{S}^{\boldsymbol{Y}^D}$
$\mathbb{F}^{\mathsf{Y}^d}$
Y^d
δ_d
$\mathbb{S}^{\mathsf{Y}_S|\boldsymbol{Y}^D\mathsf{D}}$
Y_S
=
$\mathbb{S}^{\mathsf{Y}^d}$
$\mathbb{F}^{\mathsf{Y}^d}$
Y^d
δ_d
*
Y_S

(95)

My intuition is that $\mathsf{D} = d$ is like a switch that "chooses" the d-th wire of $\boldsymbol{Y}^D$, but this isn't visually clear in 95

Equivalently, for all $d \in D$, $\mathbf{y}^D \in Y^D$:

$$\mathbb{S}^{\mathsf{Y}_S|\boldsymbol{Y}^D\mathsf{D}}_{\mathbf{y}^D,d} = \delta_{y^d} \tag{96}$$

where y^d is the d-th element of $\mathbf{y}^D$.

$d \mapsto \sum_{i\in D}[\![d = i]\!]\delta_{\mathsf{Y}^i}$ is a version of $\mathbb{S}^{\mathsf{Y}_S|\boldsymbol{Y}^D\mathsf{D}}$.

Finally, we need to account for observed variables. Let the original random variable Y be an "observed variable". Suppose further that there exists at least one additional observed variable X taking values in D which we will call the *suppositional subject* (X represents "the feature of the world that is change by supposing"). Recall that by supposition stability, both Y and X are independent of D (whatever I think might result from supposing d doesn't change what I've actually seen).

Potential outcomes features a standard assumption of *consistency*. Given random variable Y, potential outcomes $\boldsymbol{Y}^D$ and suppositional subject X, a suppositional map $\mathbb{S}$ obeys consistency iff for all $y^D \in Y^D$, $d \in D$:

$$\mathbb{S}^{\mathsf{Y}|\boldsymbol{Y}^D\mathsf{X}}_{\mathbf{y}^D,d} = \delta_{y^d} \tag{97}$$

or, as it is more typically stated

$$\mathsf{X} = d' \implies \mathsf{Y} = \mathsf{Y}^{d'} \tag{98}$$

Definition 98 is informal in our framework; a more precise statement is that conditioning on $\mathbb{1}_{\{d'\}}(\mathsf{X})$ leads to $\mathsf{Y} = \mathsf{Y}^{d'}$ almost surely. However, this is also informal as we haven't defined conditioning, which is not the same thing as conditional probability. In any case, equation 97 is well-defined, captures the core idea of consistency and doesn't require additional machinery to make sense of.

A consequence of 97 is

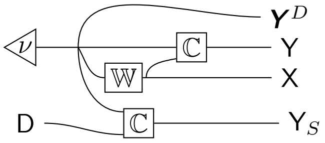
$$\mathbb{S}^{\mathsf{Y}|\boldsymbol{\mathsf{Y}}^D\mathsf{X}} = \mathbb{S}^{\mathsf{Y}_S|\boldsymbol{\mathsf{Y}}^D\mathsf{D}} \tag{99}$$

Which captures the key idea of the consistency assumption: *if what we suppose is something that actually happened, then the consequences of that supposition are the same as whatever actually happened.* This formalisation requires via X and X_S the identification of suppositions with an observed random variable. The existence of such a pair X and X_S appears to be related to the *Stable Unit Treatment Value Assumption* (SUTVA), a basic assumption of the Potential Outcomes approach. SUTVA is actually two assumptions, one of which is informally stated as "there is only one version of the treatment" (Rubin, 2005). Suppose we had a pair of variables X and X_S such that $\mathbb{S}_d^{\mathsf{X}_S|\mathsf{D}} = \delta_{f(d)}$ for some non-invertible f. Then a value of X_S cannot be uniquely associated with a supposition in D.

The existence of some X_S that D-causes Y_S and can be identified with X seems to be all we need for consistency, which is weaker than the assumptions given above

In any case, defining $\mathbb{C} := \mathbb{S}^{\mathsf{Y}_S|\boldsymbol{\mathsf{Y}}^D\mathsf{D}}$, $\mathbb{W} := \mathbb{S}^{\mathsf{X}|\boldsymbol{\mathsf{Y}}^D}$ and $\nu := \mathbb{S}^{\boldsymbol{\mathsf{Y}}^D}$, given 99 the supposition map $\mathbb{S}$ can be represented in the following form:

$\mathbb{S} =$



(100)

So far, I've argued that given a supposition map where *supposition stability* holds for a set of variables $\{\boldsymbol{\mathsf{Y}}^D, \mathsf{Y}, \mathsf{X}\}$, $\boldsymbol{\mathsf{Y}}^D$ act as *potential outcomes* for Y as in 95 and $\mathbb{S}$ obeys *consistency* with respect to suppositional subject X, potential outcomes $\boldsymbol{\mathsf{Y}}^D$ and Y, then it can be represented as in 100.

Suppose that for some supposition map $\mathbb{S} : D \to \Delta(\mathcal{Y}^D \otimes \mathcal{Y} \otimes \mathcal{X} \otimes \mathcal{Y}$, there exists some $\nu \in \Delta(\mathcal{Y}^D)$, $\mathbb{C} \in D \times Y^D \to \Delta(\mathcal{Y})$ and $\mathbb{W} : Y^D \to \Delta(\mathcal{D})$

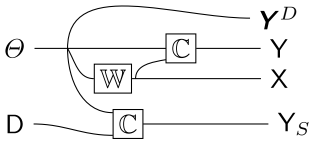
such that 100 holds where we make the obvious choices of random variable defintions. Then is is straightforward to show that consistency and supposition stability hold, but in general $\mathbf{Y}^D$ are not potential outcomes for Y. For a simple example, suppose that there exists some $d \in D$, $\mathbf{y}^D \in Y^D$, $A \in \mathcal{Y}$ such that $0 < \mathbb{C}_{d,\mathbf{y}^D}(A) < 1$, i.e. $\mathbb{C}$ is nondeterministic. Then $\mathbb{C}$ cannot obey Equation 96, which requires $\mathbb{C}$ to be deterministic.

This should be a theorem, I just didn't realise that's what I was writing when I started off.

If $\mathbb{C}$ is deterministic, then it might be possible to find a set of potential outcomes for Y, but these are not necessarily $\mathbf{Y}^D$

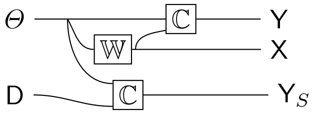
5.1 Supposition maps to see-do models

The supposition map in Equation 100 corresponds to a potential outcomes model with a known distribution of potential outcomes. If we identify the hypothesis space Θ with the space of potential outcomes Y^D, the model 100 could be considered to be the product of a prior ν on the hypothesis space and the see-do model:



(101)

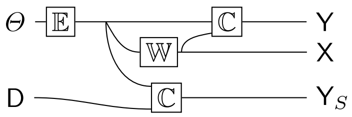
Note that 102 doesn't quite represent the inference problem, as random variables $\mathbf{Y}^D$ are actually unobserved. The inference problem is instead represented by the see-do model that marginalises over this variable:



(102)

Suppose the random variables discussed so far are sequences - i.e. $\mathbf{Y}^D = \underline{\bigotimes}_{i\in[N]}\mathbf{Y}_i^D$ and similarly for $\mathsf{X}, \mathsf{Y}, \mathsf{D}$. We might want to assume that any "legitimate" distribution ν makes $[\mathbf{Y}_1^D, \mathbf{Y}_2^D, ...]$ independent and identically distributed (see Rubin (2005) for the assumption that given a prior, the given sequence is exchangeable).

In general, given some class of allowable distributions over potential outcomes, an alternative choice is to consider the hypothesis space Θ to consist of the set of allowable distributions $\Theta \subset \Delta(\mathcal{Y}^D)$ and, letting $\mathbb{E} : \Theta \to \Delta(\mathcal{Y}^D)$ be the evaluation map $\mathbb{E}_\theta(A) = \theta(A)$, the modified see-do model becomes



(103)

Model 103 is *not* a see-do model: in a see-do model, the consequences of an act are independent of the observations conditioned on the state Θ, while 103 does not in general have this property.

6 Ignorability Does Not Identify the Average Causal Effect

I think this bit is not super interesting. The only potential point of interest is that it defuses the claim that the assumptions above are stronger than the assumptions of Potential Outcomes, as Potential Outcomes needs some more assumptions to work.

There is an error in derivations of the identification of the Average Causal Effect from the assumption of ignorability in Potential Outcomes. I will present a counterexample to the claim that ignorability + positivity leads to the identification of Average Causal Effect. The problem arises because authors assume that the average of an arbitrary sequence of random variables - absent any assumptions of IID or exchangeability - converges to something meaningful. I think the actual condition they want is “treatment exchangeability”.

For the setup, I refer to Angrist and Pischke (2014)

(although I should probably actually refer to Rubin (2005)).

> We use the letter Y as shorthand for health, the outcome variable of interest. To make it clear when were talking about specific people, we use subscripts as a stand-in for names: Y_i is the health of individual i. The outcome Y_i is recorded in our data. But, facing the choice of whether to pay for health insurance, person i has two potential outcomes, only one of which is observed. To distinguish one potential outcome from another, we add a second subscript: The road taken without health insurance leads to Y_{0i} (read this as y-zero-i) for person i, while the road with health insurance leads to Y_{1i} (read this as y-onei) for person i. Potential outcomes lie at the end of each road one might take. The causal effect of insurance on health is the difference between them, written $\mathsf{Y}_{1i} - \mathsf{Y}_{0i}$.
>
> [...]
>
> Actual health outcomes: Y_i, treatment: D_i
>
> [...]

> κ is both the individual and average causal effect on health outcomes.
>
> [...]
>
> $$\mathrm{Avg}_n[\mathsf{Y}_{1i} - \mathsf{Y}_{0i}] = \frac{1}{n}\sum_{i=1}^{n}[\mathsf{Y}_{1i} - \mathsf{Y}_{0i}] \tag{104}$$
>
> [...]
>
> $$\begin{aligned}\mathrm{Avg}_n[\mathsf{Y}_{1i}|\mathsf{D}_i = 1] - \mathrm{Avg}_n[\mathsf{Y}_{0i}|\mathsf{D}_i = 0] = \\ \kappa + \mathrm{Avg}_n[\mathsf{Y}_{0i}|\mathsf{D}_i = 1] - \mathrm{Avg}_n[\mathsf{Y}_{0i}|\mathsf{D}_i = 0]\end{aligned} \tag{105}$$
>
> [...]
>
> **Random assignment eliminates selection bias** When D_i is randomly assigned, $\mathbb{E}[\mathsf{Y}_{0i}|\mathsf{D}_i = 1] = \mathbb{E}[\mathsf{Y}_{0i}|\mathsf{D}_i = 0]$, and the difference in expectations by treatment status captures the causal effect of treatment:
>
> $$\begin{aligned}\mathbb{E}[\mathsf{Y}_i|\mathsf{D}_i = 1] - \mathbb{E}[\mathsf{Y}_i|\mathsf{D}_i = 0] &= \mathbb{E}[\mathsf{Y}_{1i}|\mathsf{D}_i = 1] - \mathbb{E}[\mathsf{Y}_{0i}|\mathsf{D}_i = 0] && (106)\\ &= \mathbb{E}[\mathsf{Y}_{0i} + \kappa|\mathsf{D}_i = 1] - \mathbb{E}[\mathsf{Y}_{0i}|\mathsf{D}_i = 0] && (107)\\ &= \kappa + \mathbb{E}[\mathsf{Y}_{0i}|\mathsf{D}_i = 1] - \mathbb{E}[\mathsf{Y}_{0i}|\mathsf{D}_i = 0] && (108)\\ &= \kappa && (109)\end{aligned}$$
>
> Provided the sample size is large enough for the law of large numbers to work its magic (so we can replace the conditional averages in equation 105 with conditional expectations), selection bias disappears in a randomized experiment

The problem is that the expectations in Equation 109 *cannot* be replaced with conditional averages as defined in 104, even in the infinite limit. From the strong law of large numbers we can deduce that, given IID variables $(\mathsf{Y}_{i0}^{j}, \mathsf{D}_i^{j}) \sim \mathbb{P}(\mathsf{Y}_{i0}, \mathsf{D}_i)$ for $j \in \mathbb{N}$,

$$\lim_{n\to\infty}\sum_{j}^{n}\frac{\mathsf{Y}_{i0}^{j}[\![\mathsf{D}_i^{j} = 1]\!]}{\sum_j^n [\![\mathsf{D}_i^{j} = 1]\!]} \overset{\mathbb{P}-a.s.}{=} \mathbb{E}[\mathsf{Y}_{1i}|\mathsf{D}_i = 1] \tag{110}$$

Note that Angrist and Pischke do *not* assume that $(\mathsf{Y}_{i0}^{j}, \mathsf{D}_i^{j})$ are given - in their conventions, this would refer to repeated samples of the "same individual".

The quantity given by their "conditional average" is an average of random variables that share similar names, but are otherwise unrelated:

$$\sum_i^n \frac{\mathsf{Y}_{i0}[\![\mathsf{D}_i = 1]\!]}{\sum_i^n [\![\mathsf{D}_i = 1]\!]} \tag{111}$$

This counterexample satisfies the stronger assumptions presented in Rubin 2005, hence it doesn't quite line up with the assumptions from Angrist and Pischke; need to incorporate Rubin.

Suppose we have random variables $(\boldsymbol{\mathsf{D}}, \boldsymbol{\mathsf{Y}}_0, \boldsymbol{\mathsf{Y}}_1) := ([\mathsf{D}_0, \mathsf{D}_1, ...], [\mathsf{Y}_{00}, \mathsf{Y}_{10}, ...], [\mathsf{Y}_{01}, \mathsf{Y}_{11}, ...]) \in [0,1]^{3\mathbb{N}}$ and

$$\mathbb{P}(\boldsymbol{\mathsf{D}} = \mathbf{d}, \boldsymbol{\mathsf{Y}}_0 = \mathbf{y}_0, \boldsymbol{\mathsf{Y}}_1 = \mathbf{y}_1) = \prod_{i\in\mathbb{N}} \left((1-\epsilon)\delta_{(i \mod 2)}(d_i) + \epsilon\right) \delta_{(i \mod 2)}(y_{i0}) \delta_{(1-i \mod 2)}(y_{i1}) \tag{112}$$

By construction $\boldsymbol{\mathsf{Y}}_1, \boldsymbol{\mathsf{Y}}_0 \perp\!\!\!\perp_{\mathbb{P}} \boldsymbol{\mathsf{D}}$, and $\mathbb{P}(\mathsf{D}_i = d_i) > 0$ for all d_i, which implies for all i $\mathbb{E}[\mathsf{Y}_{0i}|\mathsf{D}_i = 1] = \mathbb{E}[\mathsf{Y}_{0i}|\mathsf{D}_i = 0]$. However

$$\lim_{n\to\infty} \sum_i^n \frac{\mathsf{Y}_{i1}[\![\mathsf{D}_i = 1]\!]}{\sum_i^n [\![\mathsf{D}_i = 1]\!]} - \lim_{n\to\infty} \sum_i^n \frac{\mathsf{Y}_{i0}[\![\mathsf{D}_i = 0]\!]}{\sum_i^n [\![\mathsf{D}_i = 0]\!]} = 1 - \frac{\epsilon}{2} - \frac{\epsilon}{2} \tag{113}$$

$$= 1 - \epsilon \tag{114}$$

$$\lim_{n\to\infty} \sum_i^n \frac{\mathsf{Y}_{i0}}{n} - \lim_{n\to\infty} \sum_i^n \frac{\mathsf{Y}_{i1}}{n} = \frac{1}{2} - \frac{1}{2} \tag{115}$$

$$= 0 \tag{116}$$

$$= \text{“the average causal effect”} \tag{117}$$

$$\neq \lim_{n\to\infty} \sum_i^n \frac{\mathsf{Y}_{i1}[\![\mathsf{D}_i = 1]\!]}{\sum_i^n [\![\mathsf{D}_i = 1]\!]} - \lim_{n\to\infty} \sum_i^n \frac{\mathsf{Y}_{i0}[\![\mathsf{D}_i = 0]\!]}{\sum_i^n [\![\mathsf{D}_i = 0]\!]} \tag{118}$$

Contradicting the claim made by Eq. 109.

7 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.

7.1 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
Blackboard $\mathbb{P}$, μ, ν:	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 $\mathcal{E}$
Blackboard capitals: $\mathbb{A}$	Markov kernel $\mathbb{A} : X \times \mathcal{Y} \to [0, 1]$ (stochastic maps
Subscripted Markov kernels: $\mathbb{A}_x$	The probability measure given by the curried Markov kern
$A \to \Delta(\mathcal{B})$	Markov kernel signature, treated as equivalent to $A \times \mathcal{B}$
$\mathbb{A} : x \mapsto \nu$	Markov kernel definition, equivalent to $\mathbb{A}(x, B) = \nu(B)$ f
Sans serif capitals: A, X	Measurable functions; we will also call them random va
$\mathbb{F}^{\mathsf{X}}$	The Markov kernel associated with the function X: $\mathbb{F}^{\mathsf{X}} \equiv$
$\mathbb{N}_{\mathsf{A}\vert\mathsf{B}}$	The conditional probability (disintegration) of A given B
$\nu\mathbb{F}^{\mathsf{X}}$	The marginal distribution of X under ν

7.2 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 $\mathbb{M} : E \times \mathcal{F} \to [0,1]$ such that

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

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

The map $x \mapsto \mathbb{M}_x$ is of type $E \to \Delta(\mathcal{F})$. We will abuse notation somewhat to write $\mathbb{M} : E \to \Delta(\mathcal{F})$, 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 $\mathbb{N} : \{*\} \to \Delta(\mathcal{E})$ with the probability measure $\mathbb{N}_*$. In addition, there is a unique Markov kernel $\ast : E \to \Delta(\{\{*\}, \emptyset\})$ given by $x \mapsto \delta_*$ for all $x \in E$ which we will call the "discard" map.

7.3 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 $\mathbb{M} : X \to \Delta(\mathcal{Y})$ and $\mathbb{N} : Y \to \Delta(\mathcal{Z})$, the measure-kernel product $\mu\mathbb{A}(G) := \int \mathbb{A}_x(G) d\mu(x)$ yields a probability measure $\mu\mathbb{A}$ on $\mathcal{Z}$, the kernel-kernel product $\mathbb{M}\mathbb{N}(x; H) = \int_Y \mathbb{B}(y; H) d\mathbb{A}_x$ yields a kernel $\mathbb{M}\mathbb{N} : X \to \Delta(\mathcal{Z})$ and the kernel-function product $\mathbb{A}\mathsf{T}(x) := \int_Y \mathsf{T}(y) d\mathbb{A}_x$ yields a measurable function $X \to T$. Kernel products are associative (Çinlar, 2011).

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

7.4 String Diagrams

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

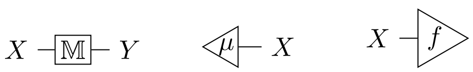
Such constructions can, however, be straightforwardly captured with string diagrams, a notation developed for category theoretic probability. Cho and Jacobs (2019) also provides an extensive introduction to the notation discussed here.

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 $\mathbb{M} : X \to \Delta(\mathcal{Y})$ 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 **??**).

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

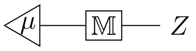


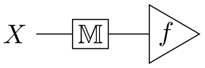
(119)

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

$$\mathbb{M}\mathbb{N} := X \text{ --- }\boxed{\mathbb{M}}\text{ --- }\boxed{\mathbb{N}}\text{ --- } Z \tag{120}$$

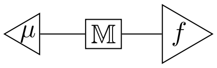
Probability measures are distinguished in that that they only admit "right composition" while functions only admit "left composition". For $\mu \in \Delta(\mathcal{E})$, $h : F \to X$:

$\mu\mathbb{M} :=$  (121)

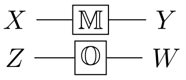
$\mathbb{M}f :=$  (122)

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

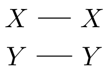
$$\mathbb{E}_{\mu\mathbb{M}}(f) = \mu\mathbb{M}f \tag{123}$$

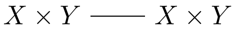
$:=$  (124)

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

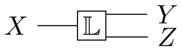
$\mathbb{M} \otimes \mathbb{N} :=$  (125)

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

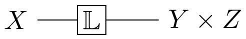
$X \times Y \cong \mathrm{Id}_X \otimes \mathrm{Id}_Y :=$  (126)

$=$  (127)

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

 (128)

$\equiv$ (129)

 (130)

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

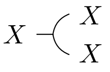
$$\mathbf{Id} : x \mapsto \delta_x \equiv \; X \text{ --- } X \tag{131}$$

The identity kernel acts as the identity under left or right products:

$$\begin{aligned}
(\mathbb{K}\mathbf{Id})_w(A) &= \int_X \mathbf{Id}_x(A) d\mathbb{K}_w(x) && (132)\\
&= \int_X \delta_x(A) d\mathbb{K}_w(x) && (133)\\
&= \int_A d\mathbb{K}_w(x) && (134)\\
&= \mathbb{K}_w(A) && (135)\\
(\mathbf{Id}\mathbb{K})_w(A) &= \int_X \mathbb{K}_x(A) d\mathbf{Id}_w(x) && (136)\\
&= \int_X \mathbb{K}_x(A) d\delta_w(x) && (137)\\
&= \mathbb{K}_w(A) && (138)
\end{aligned}$$

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

$$\curlyvee : x \mapsto \delta_{(x,x)} \equiv \tag{139}$$



The copy map "copies" its arguments under an integral:

$$\int_{(} X \times X) f(x, x', x'') d\curlyvee_x(x', x'') = \int_{(} X \times X) f(x, x', x'') d\delta_{(x,x)}(x', x'') \tag{140}$$

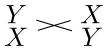
$$= f(x, x, x) \tag{141}$$

$$\int_W \int_{(} X \times X) f(x', x'') d\curlyvee_w(x', x'') d\mu(w) \tag{142}$$

$$= \int_W f(w, w) d\mu(w) \tag{143}$$

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

$$\sigma := (x, y) \to \delta_{(y,x)} \equiv \tag{144}$$



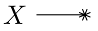
The swap map swaps its arguments under an integral:

$$\int_{(} X \times X) f(x, x') d\sigma_{(x_0,x_1)}(x, x') = \int_{(} X \times X) f(x, x') d\delta_{(x_1,x_0)}(x, x') \tag{145}$$

$$= f(x_1, x_0) \tag{146}$$

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

$$\ast : x \mapsto \delta_* \equiv \tag{147}$$



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 124), the product of a probability measure with the associated Markov kernel is the pushforward measure $f_{\#}\mu$.

Definition 7.1 (Function induced kernel)**.** Given a measurable function $g : F \to X$, define the function induced kernel $\mathbb{F}^g : F \to \Delta(\mathcal{X})$ to be the the Markov kernel $a \mapsto \delta_{g(a)}$ for all $a \in X$.

Definition 7.2 (Pushforward kernel)**.** Given a kernel $\mathbb{M} : E \to \Delta(\mathcal{F})$ and a measurable function $g : F \to X$, the *pushforward kernel* $g_{\#}\mathbb{M} : E \to \Delta(\mathcal{X})$ is the kernel such that $g_{\#}\mathbb{M}(a; B) = \mathbb{M}(a; g^{-1}(B))$.

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

Lemma 7.3 (Pushforward kernels are functional kernel products)**.** *Given a kernel $\mathbb{M} : E \to \Delta(\mathcal{F})$ and a measurable function $g : F \to X$,the pushforward $g_{\#}\mathbb{M} = \mathbb{M}\mathbb{F}^g$.*

Proof.

$$\mathbb{M}\mathbb{F}^g(a; B) = \int_F \delta_{g(y)}(B) d\mathbb{M}_a(y) \tag{148}$$

$$= \int_F \delta_y(g^{-1}(B)) d\mathbb{M}_a(y) \tag{149}$$

$$= \int_{g^{-1}(B)} d\mathbb{M}_a(y) \tag{150}$$

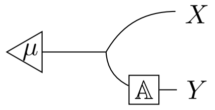
$$= g_{\#}\mathbb{M}(a; B) \tag{151}$$

□

7.4.1 Comparison of notations

We are in a position to compare the three introduced notations using a few examples. Given $\mu \in \Delta(X), \mathbb{A} : X \to \Delta(Y)$ 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:

$$\nu := A \times B \mapsto \int_A A(x;B)d\mu(x) \equiv \mu\curlyvee(\mathbf{Id}_X \otimes \mathbb{A}) \equiv \tag{152}$$



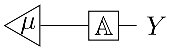
Where the resulting object is a probability measure $\nu \in \Delta(\mathcal{X}\otimes\mathcal{Y})$. Note that the elementary notation requires a function definition here, while the product and string notations can represent the measure without explicitly addressing its action on various inputs and outputs. Cho and Jacobs (2019) calls this construction "integrating $\mathbb{A}$ 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 152 over X in our three separate notations as follows:

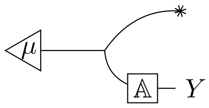
$$\nu_Y(B) = \nu(X \times B) = \int_X A(x;B)d\mu(x) \tag{153}$$

$$\nu_Y = \mu\mathbb{A} = \mu\curlyvee(\mathbf{Id}_X \otimes \mathbb{A})(\ast \otimes \mathbf{Id}_Y) \tag{154}$$

$$\nu_Y = \tag{155}$$



$=$



The elementary notation 153 makes the relationship between ν_Y and ν explicit and, again, requires the action on each event to be defined. The product notation 154 is, in my view, the least transparent but also the most compact in the form $\mu\mathbb{A}$, 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 in152, 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.

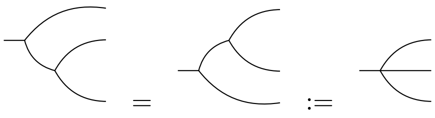
7.4.2 Working With String Diagrams

todo:

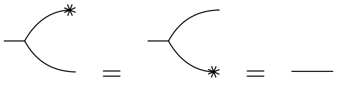
- Functional generalisation
- Conditioning
- Infinite copy map
- De Finetti's representation theorem

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.

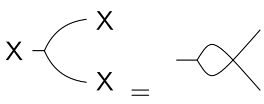
Axioms of Symmetric Monoidal Categories For the following, we either omit labels or label diagrams with their domain and codomain spaces, as we are discussing identities of kernels rather than identities of components of a condtional probability space. Recalling the unique Markov kernels defined above,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



(156)

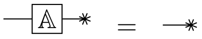


(157)



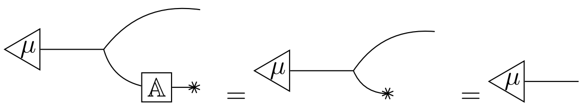
(158)

The discard map ∗ can "fall through" any Markov kernel:



(159)

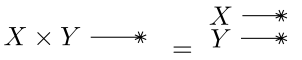
Combining 247 and 159 we can derive the following: integrating $\mathbb{A} : X \to \mathit{\Delta}(\mathcal{Y})$ with respect to $\mu \in \mathit{\Delta}(\mathcal{X})$ and then discarding the output of $\mathbb{A}$ leaves us with μ:



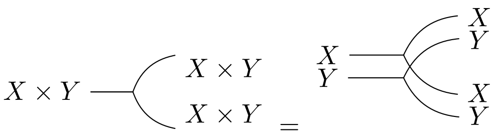
(160)

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

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

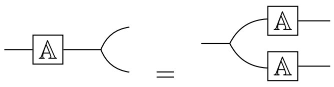


(161)



(162)

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



(163)

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

Definition 7.4 (Deterministic Markov kernel)**.** A *deterministic* Markov kernel $\mathbb{A} : E \to \Delta(\mathcal{F})$ is a kernel such that $\mathbb{A}_x(B) \in \{0,1\}$ for all $x \in E$, $B \in \mathcal{F}$.

Theorem 7.5 (Copy map commutes for deterministic kernels (Fong, 2013))**.** *Equation 163 holds iff* $\mathbb{A}$ *is deterministic.*

7.5 Random Variables

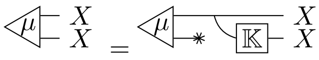
The summary of this section is:

- 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 label 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.

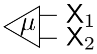
We use the same approach here, with a couple of additions. We are interested in variables whose outcomes depend both on random processes and decisions. Suppose that given a particular distribution over decision variables, a probability distribution over the decision variables and random variables is obtained. Such a model is described by a Markov kernel rather than a probability distribution. We therefore investigate *Markov kernel spaces*.

In the graphical notation that we are using, random variables can be thought of as a means of assigning unambiguous names to each wire in a set of diagrams. In order to do this, it is necessary to suppose that all diagrams in the set describe properties of an *ambient Markov kernel* or *ambient probability measure*. Consider the following example with the ambient probability measure $\mu \in \Delta(\mathcal{X}\otimes\mathcal{X})$. Suppose we have a Markov kernel $\mathbb{K} : X \to \Delta(\mathcal{X})$ such that the following holds:



(164)

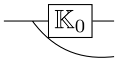
Suppose that we also assign the names X_1 to the upper output wire and X_2 to the lower output wire in the diagram of μ:



(165)

Then it seems sensible to call $\mathbb{K}$ “the probability of X_2 given X_1”. We will make this precise so that it matches the usual notion of the probability of one variable given another (see Çinlar (2011) for a definition of this usual notion).

Definition 7.6 (Probability space, Markov kernel space)**.** 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

$\mathbb{K} :=$ 

(166)

For brevity, we will omit the σ-algebras in further definitions of Markov kernel spaces: $(\mathbb{K}, \Omega, D)$.

A *probability space* $(\mathbb{P}, \Omega, \mathcal{F})$ is a probability measure $\mathbb{P} : \Delta(\Omega)$, which we call the *ambient measure*, along with the *sample space* Ω and the *events* $\mathcal{F}$. A probability space is equivalent to a Markov kernel space with domain $D = \{*\}$ - note that $\Omega \times \{*\} \cong \Omega$.

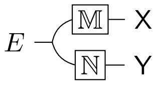
Definition 7.7 (Random variable)**.** Given a Markov kernel space $(\mathbb{K}, \Omega, D)$, a random variable X is a measurable function $\Omega \times D \to E$ for arbitrary measurable E.

Definition 7.8 (Domain variable)**.** Given a Markov kernel space $(\mathbb{K}, \Omega, D)$, the *domain variable* $\mathsf{D} : \Omega \times D \to D$ is the distinguished random variable $\mathsf{D} : (x, d) \mapsto d$.

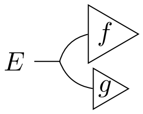
Unlike random variables on probability spaces, random variables on Markov kernel spaces do not generally have unique marginal distributions. An analogous operation of *marginalisation* can be defined, but the result is generally a Markov kernel. We will define marginalisation via coupled tensor products.

Definition 7.9 (Coupled tensor product $\underline{\otimes}$)**.** Given two Markov kernels $\mathbb{M}$ and $\mathbb{N}$ or functions f and g with shared domain E, let $\mathbb{M}\underline{\otimes}\mathbb{N} := \curlyvee(\mathbb{M} \otimes \mathbb{N})$ and $f\underline{\otimes}g := \curlyvee(f \otimes g)$ where these expressions are interpreted using standard product notation. Graphically:

$$\mathbb{M}\underline{\otimes}\mathbb{N} := \tag{167}$$



$$f\underline{\otimes}g := \tag{168}$$



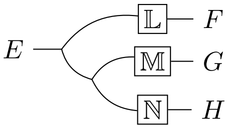
The operation denoted by $\underline{\otimes}$ is associative (Lemma 7.10), so we can without ambiguity write $f\underline{\otimes}g\underline{\otimes}h = (f\underline{\otimes}g)\underline{\otimes}h = f\underline{\otimes}(g\underline{\otimes}h)$ for finite groups of functions or Markov kernels sharing a domain.

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

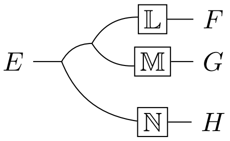
Lemma 7.10 ($\underline{\otimes}$ is associative)**.** *For Markov kernels $\mathbb{L} : E \to \delta(\mathcal{F})$, $\mathbb{M} : E \to \delta(\mathcal{G})$ and $\mathbb{N} : E \to \delta(\mathcal{H})$, $(\mathbb{L}\underline{\otimes}\mathbb{M})\underline{\otimes}\mathbb{N} = \mathbb{L}\underline{\otimes}(\mathbb{M}\underline{\otimes}\mathbb{N})$.*

Proof.

$$\mathbb{L}\underline{\otimes}(\mathbb{M}\underline{\otimes}\mathbb{N}) = \tag{169}$$



$$= \tag{170}$$



$$= (\mathbb{L}\underline{\otimes}\mathbb{M})\underline{\otimes}\mathbb{N} \tag{171}$$

This follows directly from Equation 246. □

Definition 7.11 (Marginal distribution, marginal kernel)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ and the random variable $\mathsf{X} : \Omega \to G$ the *marginal distribution* of X is the probability measure $\mathbb{P}^{\mathsf{X}} := \mathbb{P}\mathbb{F}^{\mathsf{X}}$.

See Lemma 7.3 for the proof that this matches the usual definition of marginal distribution.

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

Definition 7.12 (Joint distribution, joint kernel)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ and the random variables $\mathsf{X} : \Omega \to G$ and $\mathsf{Y} : \Omega \to H$, the *joint distribution* of X and Y, $\mathbb{P}^{\mathsf{XY}} \in \Delta(\mathcal{G} \otimes \mathcal{H})$, is the marginal distribution of $\mathsf{X}\underline{\otimes}\mathsf{Y}$. That is, $\mathbb{P}^{\mathsf{XY}} := \mathbb{P}\mathbb{F}^{\mathsf{X}\underline{\otimes}\mathsf{Y}}$

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 $\mathbb{K}\mathbb{F}^{\mathsf{X}\underline{\otimes}\mathsf{Y}}$.

Joint distributions and kernels have a nice visual representation, as a result of Lemma 7.13 which follows.

Lemma 7.13 (Product marginalisation interchange)**.** *Given two functions, the kernel associated with their coupled product is equal to the coupled product of the kernels associated with each function.*

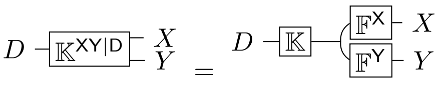
Given $\mathsf{X} : \Omega \to G$ *and* $\mathsf{Y} : \Omega \to H$*,* $\mathbb{F}^{\mathsf{X}\underline{\otimes}\mathsf{Y}} = \mathbb{F}^{\mathsf{X}}\underline{\otimes}\mathbb{F}^{\mathsf{Y}}$

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

$$\begin{aligned}
\mathbb{F}^{\mathsf{X}\underline{\otimes}\mathsf{Y}}(a; B \times C) &= \delta_{\mathsf{X}(a),\mathsf{Y}(a)}(B \times C) && (172)\\
&= \delta_{\mathsf{X}(a)}(B)\delta_{\mathsf{Y}(a)}(C) && (173)\\
&= (\delta_{\mathsf{X}(a)} \otimes \delta_{\mathsf{Y}(a)})(B \times C) && (174)\\
&= \mathbb{F}^{\mathsf{X}}\underline{\otimes}\mathbb{F}^{\mathsf{Y}} && (175)
\end{aligned}$$

Equality follows from the monotone class theorem. □

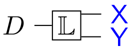
Corollary 7.14. *Given a Markov kernel space* $(\mathbb{K}, \Omega, D)$ *and random variables* $\mathsf{X} : \Omega \times D \to X$*,* $\mathsf{Y} : \Omega \times D \to Y$*, the following holds:*


D
$\mathbb{K}^{\mathsf{XY}|\mathsf{D}}$
X
Y
=
D
$\mathbb{K}$
$\mathbb{F}^{\mathsf{X}}$
X
$\mathbb{F}^{\mathsf{Y}}$
Y

(176)

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

Definition 7.15 (Wire labels - joint kernels)**.** Suppose we have a Markov kernel space $(\mathbb{K}, \Omega, D)$, random variables $\mathsf{X} : \Omega \times D \to X$, $\mathsf{Y} : \Omega \times D \to Y$ and a Markov kernel $\mathbb{L} : D \to \Delta(\mathcal{X} \times \mathcal{Y})$. The following *output labelling* of $\mathbf{L}$:

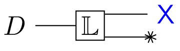


(177)

is *valid* iff

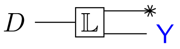
$$\mathbb{L} = \mathbb{K}_{\mathsf{XY|D}} \tag{178}$$

and



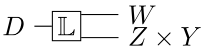
$$= \mathbb{K}^{\mathsf{X|D}} \tag{179}$$

and



$$= \mathbb{K}^{\mathsf{Y|D}} \tag{180}$$

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

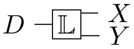


(181)

For *this* diagram, properties 179 and 180 do not hold, even though 178 does.

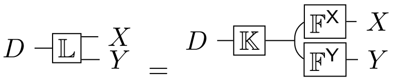
Lemma 7.16 (Output label assignments exist)**.** *Given Markov kernel space* $(\mathbb{K}, \Omega, D)$*, random variables* $\mathsf{X} : \Omega \times D \to X$ *and* $\mathsf{Y} : \Omega \times D \to Y$ *then there exists a diagram of* $\mathbb{L} := \mathbb{K}^{\mathsf{XY|D}}$ *with a valid output labelling assigning* X *and* Y *to the output wires.*

Proof. By definition, $\mathbb{L}$ has signature $D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$. Thus, by the rule that tensor product spaces can be represented by parallel wires, we can draw



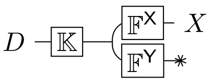
(182)

By Corollary 7.14, we have



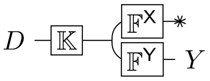
(183)

Therefore



$$= \mathbb{K}\mathbb{F}^{\mathsf{X}} \tag{184}$$

$$= \mathbb{K}^{\mathsf{X|D}} \tag{185}$$



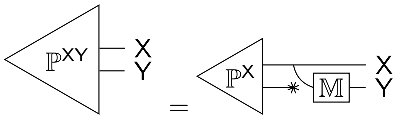
$$= \mathbb{K}\mathbb{F}^{\mathsf{Y}} \qquad (186)$$

$$= \mathbb{K}^{\mathsf{Y|D}} \qquad (187)$$

□

In all further work, wire labels will be used without special colouring.

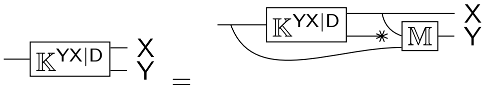
Definition 7.17 (Disintegration)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$, and random variables X and Y, we say that $\mathbb{M} : E \to \Delta(\mathcal{F})$ is a Y *on* X *disintegration* of $\mathbb{P}$ iff



(188)

$\mathbb{M}$ is a version of $\mathbb{P}^{\mathsf{Y|X}}$, "the probability of Y given X". Let $\mathbb{P}^{\{\mathsf{Y|X}\}}$ be the set of all kernels that satisfy 188 and $\mathbb{P}^{\mathsf{Y|X}}$ an arbitrary member of $\mathbb{P}^{\mathsf{Y|X}}$.

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



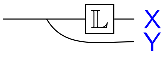
(189)

Write $\mathbb{K}^{\{\mathsf{Y|XD}\}}$ for the set of kernels satisfying 189 and $\mathbb{K}^{\mathsf{Y|XD}}$ for an arbitrary member of $\mathbb{K}^{\{\mathsf{Y|XD}\}}$.

Definition 7.18 (Wire labels – input)**.** An input wire is *connected* to an output wire if it is possible to trace a path from the start of the input wire to the end of the output wire without passing through any boxes, erase maps or right facing triangles.

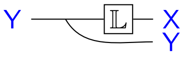
If an input wire is connected to an output wire and that output wire has a valid label X, then it is valid to label the input wire with X.

For example, if the following are valid output labels with respect to $(\mathbb{P}, \Omega)$:



(190)

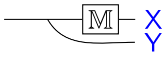
i.e. if $\mathbb{L} \in \mathbb{P}^{\{\mathsf{XY|Y}\}}$, then the following is a valid input label:



(191)

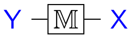
An input wire in a diagram for $\mathbb{M}$ may be labeled X *if and only if* copy and identity maps can be inserted to yield a diagram in which the input wire labeled X is connected to an output wire with valid label X.

So, if $\mathbb{M} \in \mathbb{P}^{\{\mathsf{X}|\mathsf{Y}\}}$, then it is straightforward to show that



$$\in \mathbb{P}^{\{\mathsf{XY}|\mathsf{Y}\}} \qquad (192)$$

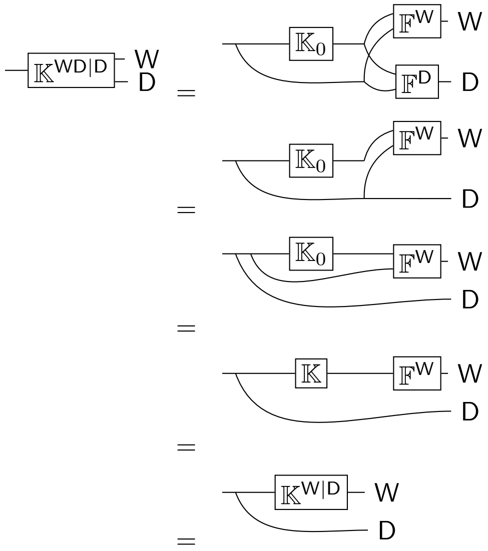
and hence the output labels are valid. Diagram 192 is constructed by taking the product of the copy map with $\mathbb{M} \otimes \mathbf{Id}$. Thus it is valid to label $\mathbb{M}$ with



(193)

Lemma 7.19 (Labeling of disintegrations)**.** *Given a kernel space* $(\mathbb{K}, \Omega, D)$*, random variables* X *and* Y*, domain variable* D *and disintegration* $\mathbb{L} \in \mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$*, there is a diagram of* $\mathbb{L}$ *with valid input labels* X *and* D *and valid output label* Y*.*

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



(194)

(195)

(196)

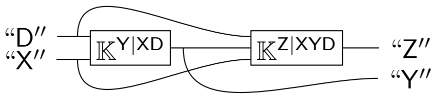
(197)

(198)

□

We also use the informal convention of labelling wires in quote marks "X'' if that wire is "supposed to" carry the label X but the label may not be valid.

Theorem 7.20 (Iterated disintegration)**.** *Given a kernel space* $(\mathbb{K}, \Omega, D)$*, random variables* X*,* Y *and* Z *and domain variable* D*,*


"D"
"X"
$\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}$
$\mathbb{K}^{\mathsf{Z}|\mathsf{XYD}}$
"Z"
"Y"

$$\in \mathbb{K}^{\{\mathsf{ZY}|\mathsf{XD}\}} \tag{199}$$

Equivalently, for $d \in D$ *and* $x \in X$*,* $A \in \mathcal{Y}$*,* $B \in \mathcal{Z}$*,*

$$(d, x; A, B) \mapsto \int_A \mathbb{K}^{\mathsf{Z}|\mathsf{XYD}}_{(x,y,d)}(B) d\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}_{(x,d)}(y) \in \mathbb{K}^{\{\mathsf{ZY}|\mathsf{XD}\}} \tag{200}$$

Proof.

write this up

□

The existence of disintegrations of standard measurable probability spaces is well known.

Theorem 7.21 (Disintegration existence - probability space)**.** *Given a probability measure* $\mu \in \Delta(\mathcal{X} \otimes \mathcal{Y})$*, if* $(F, \mathcal{F})$ *is standard then a disintegration* $\mathbb{K} : X \to \Delta(\mathcal{Y})$ *exists (Çinlar, 2011).*

In particular, if for all $x \in X$, $\mathbb{P}^{\mathsf{X}}(\mathsf{X} \in \{x\}) > 0$, then $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_x(\mathsf{Y} \in A) = \frac{\mathbb{P}^{\mathsf{XY}}(\mathsf{Y} \in A \ \&\ \mathsf{X} \in \{x\})}{\mathbb{P}^{\mathsf{X}}(\mathsf{X} \in \{x\})}$.

For Markov kernel spaces, we make the simplifying assumption that the domain space D is a discrete space. Given this assumption, there exists a positive definite probability $\mu \in \Delta(\mathcal{D})$. That is, for every $d \in D$, $\mu(\{d\}) > 0$. Given this assumption, for every Markov kernel space $(\mathbb{K}, \Omega, D)$ there is a probability space $(\mathbb{P}, \Omega \times D)$ such that $\mathbb{K}$ can be uniquely defined as a disintegration of $\mathbb{P}$. For uncountable D, even if it is standard measurable, this is not possible (Hájek, 2003).

Definition 7.22 (Relative probability space)**.**

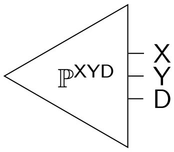
better name

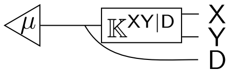
Given a Markov kernel space $(\mathbb{K}, \Omega, D)$ and a positive definite measure $\mu \in \Delta(\mathcal{D})$, $(\mu\mathbb{K}, \Omega \times D)$ is a *relative* probability space.

For any random variable $\mathsf{X} : \Omega \times D \to X$ on $(\mathbb{K}, \Omega, D)$, its relative on $(\mu\mathbb{K}, \Omega \times D)$ is given by the same measurable function, and we give it the same name X.

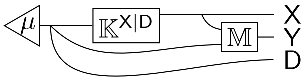
Lemma 7.23 (Agreement of disintegrations)**.** *Given a Markov kernel space* $(\mathbb{K}, \Omega, D)$*, any relative probability space* $(\mu\mathbb{K}, \Omega \times D)$ *and any random variables* $\mathsf{X} : \Omega \times D \to X$*,* $\mathsf{Y} : \Omega \times D \to Y$*,* $\mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}} = (\mu\mathbb{K})^{\{\mathsf{Y}|\mathsf{XD}\}}$ *(note that this set equality).*

Proof. Define $\mathbb{P} := \mu\mathbb{K}$ and let $\mathbb{M}$ be an arbitrary version of $\mathbb{K}^{\{\mathsf{Y|XD}\}}$. Then

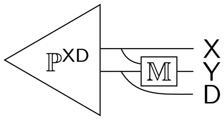




$$= \qquad (201)$$



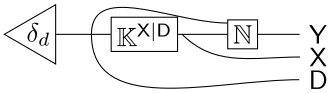
$$= \qquad (202)$$



$$= \qquad (203)$$

Thus $\mathbb{M} \in \mathbb{P}^{\{\mathsf{Y|XD}\}}$.

Let $\mathbb{N}$ be an arbitrary version of $\mathbb{P}^{\{\mathsf{Y|XD}\}}$. To show that $\mathbb{N} \in \mathbb{K}^{\{\mathsf{Y|XD}\}}$, we will show for all $d \in D$



$$\mathbb{Q} := \qquad (204)$$

$$= \mathbb{K}_d^{\mathsf{XYD|D}} \qquad (205)$$

For $A \in \mathcal{X}$,$B \in \mathcal{Y}$, $d \in D$, we have $\mathbb{Q}(A \times B \times \emptyset) = 0 = \mathbb{K}_d^{\mathsf{XYD|D}}(A \times B \times \emptyset,$ and for $\{d\} \in \mathcal{D}$ we have $\mu(\{d\}) > 0$ so:

$$
\begin{aligned}
\mathbb{Q}(A\times B\times\{d\}) &= \int_{X^2}\int_X\int_{D^3}\mathbb{N}_{d'',x'}(A)\mathbf{Id}_{x''}(B)\mathbf{Id}_{d'''}(\{d\})d{\curlyvee}_d(d',d'',d''')d\mathbb{K}^{\mathsf{X|D}}_{d'}(x)d{\curlyvee}_x(x',x'') \tag{206}\\
&= \delta_d(\{d\})\int_X \mathbb{N}_{d,x}(A)\delta_x(B)d\mathbb{K}^{\mathsf{X|D}}_d(x) \tag{207}\\
&= \frac{1}{\mu(\{d\})}\int_{\{d\}}d\mu(d')\int_X \mathbb{N}_{d,x}(A)\delta_x(B)d\mathbb{K}^{\mathsf{X|D}}_d(x) \tag{208}\\
&= \frac{1}{\mu(\{d\})}\int_D\int_X \mathbb{N}_{d,x}(A)\delta_{d'}(\{d\})\delta_x(B)d\mathbb{K}^{\mathsf{X|D}}_d(a)d\mu(d') \tag{209}\\
&= \frac{1}{\mu(\{d\})}\int_D\int_X \mathbb{N}_{d,x}(A)\delta_{d'}(\{d\})\delta_x(B)d\mathbb{K}^{\mathsf{X|D}}_{d'}(a)d\mu(d') \tag{210}\\
&= \frac{1}{\mu(\{d\})}\mathbb{P}^{\mathsf{XYD}}(A\times B\times\{d\}) \tag{211}\\
&= \frac{1}{\mu(\{d\})}\int_D \mathbb{K}^{\mathsf{XYD|D}}_{d'}(A\times B\times\{d\})d\mu(d') \tag{212}\\
&= \frac{1}{\mu(\{d\})}\int_D \mathbb{K}_{d'}\mathsf{XY|D}(A\times B)\delta_{d'}(\{d\})d\mu(d') \tag{213}\\
&= \mathbb{K}^{\mathsf{XY|D}}_d(A\times B) \tag{214}\\
&= \mathbb{K}^{\mathsf{XY|D}}_d(A\times B)\delta_d(\{d\}) \tag{215}\\
&= \int_D \mathbb{K}^{\mathsf{XY}}_{d'}(A\times B)\delta_{d''}(\{d\})d{\curlyvee}_d(d',d'') \tag{216}\\
&= \mathbb{K}^{\mathsf{XYD|D}}_d(A\times B\times\{d\}) \tag{217}
\end{aligned}
$$

Equality follows from the monotone class theorem. Thus $\mathbb{N}\in\mathbb{K}^{\{\mathsf{Y|XD}\}}$. □

Thus any kernel conditional probability $\mathbb{K}^{\mathsf{Y|XD}}$ can equally well be considered a regular conditional probability $\mathbb{P}^{\mathsf{Y|XD}}$ for a related probability space $(\mathbb{P},\Omega\times D)$ under the obvious identification of random variables, provided D is countable. Note that any conditional probability $\mathbb{P}^{\mathsf{Y|X}}$ that is *not* conditioned on D is undefined in the kernel space $(\mathbb{K},\Omega,D)$.

7.5.1 Conditional Independence

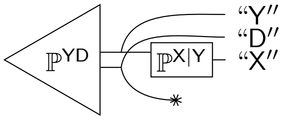
Definition 7.24 (Kernels constant in an argument)**.** Given a kernel $(\mathbb{K},\Omega,D)$ and random variables Y and X, we say a verstion of the disintegration $\mathbb{K}^{\mathsf{Y|XD}}$ is constant in D if for all $x\in X$, $d,d'\in D$, $\mathbb{K}^{\mathsf{Y|XD}}_{(x,d)}=\mathbb{K}^{\mathsf{Y|XD}}_{(x,d')}$.

Definition 7.25 (Domain Conditional Independence)**.** Given a kernel space $(\mathbb{K},\Omega,D)$, relative probability space $(\mathbb{P},\Omega\times D)$, variables X,Y and domain variable D, X is *conditionally independent* of D given Y, written $\mathsf{X}\perp\!\!\!\perp_{\mathbb{K}}\mathsf{D}|\mathsf{Y}$ if any of the following equivalent conditions hold:

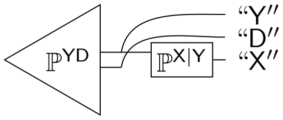
Almost sure equality

- $\mathbb{P}^{\mathsf{XD|Y}} \sim \mathbb{P}^{\mathsf{X|Y}} \underline{\otimes} \mathbb{P}^{\mathsf{D|Y}}$
- For any version of $\mathbb{P}^{\{\mathsf{X|Y}\}}$, $\mathbb{P}^{\mathsf{X|Y}} \otimes *_D$ is a version of $\mathbb{K}^{\{\mathsf{X|YD}\}}$
- There exists a version of $\mathbb{K}^{\{\mathsf{X|YD}\}}$ constant in D

Theorem 7.26 (Definitions are equivalent)**.** *(1) $\implies$ (2): By Lemma 7.23, $\mathbb{P}^{\{\mathsf{Y|XD}\}} = \mathbb{K}^{\{\mathsf{Y|XD}\}}$. Thus it is sufficient to show that $\mathbb{P}^{\mathsf{X|Y}} \otimes *$ is a version of $\mathbb{P}^{\{\mathsf{X|YD}\}}$.*

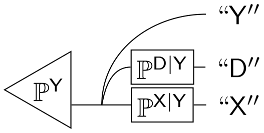


$=$



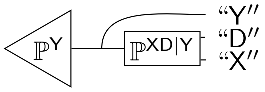
(218)

$=$


$\mathbb{P}^{\mathsf{Y}}$
$\mathbb{P}^{\mathsf{D|Y}}$
$\mathbb{P}^{\mathsf{X|Y}}$
"Y"
"D"
"X"

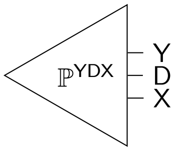
(219)

$\overset{condition(1)}{=}$



(220)

$=$



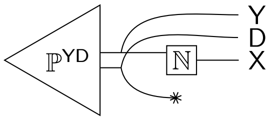
(221)

(2) $\implies$ (3)

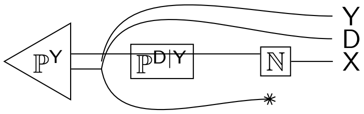
$\mathbb{P}^{\mathsf{X|Y}} \otimes *_D$ is a version of $\mathbb{K}^{\{\mathsf{X|YD}\}}$ by assumption, and is clearly constant in D.

(3) $\implies$ (1)

By lemma 7.23, there also exists a version of $\mathbb{P}^{\{\mathsf{X|YD}\}}$ constant in D. *Let $\mathbb{M} : Y \times D \to \Delta(\mathcal{X})$ be such a version. For arbitrary $d_0 \in D$, let $\mathbb{N} := \mathbb{M}_{(\cdot,d_0)} : Y \to \Delta(\mathcal{X})$ be the map $x \mapsto \mathbb{M}_{(x,d_0)}$. By constancy in* D, *$\mathbb{M} = * \otimes \mathbb{N}$. We wish to show $\mathbb{P}^{\mathsf{X|Y}} \underline{\otimes} \mathbb{P}^{\mathsf{D|Y}} \in \mathbb{P}^{\{\mathsf{XD|Y}\}}$. By Theorem 7.20, we have*



$=$



(222)

Definition 7.27 (Conditional probability existence)**.** Given a kernel space $(\mathbb{K}, \Omega, D)$ and random variables X, Y, we say $\mathbb{K}^{\{\mathsf{Y}|\mathsf{X}\}}$ *exists* if $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{K}} \mathsf{D}|\mathsf{X}$. If $\mathbb{K}^{\{\mathsf{Y}|\mathsf{X}\}}$ exists then it is by definition equal to $\mathbb{P}^{\{\mathsf{Y}|\mathsf{X}\}}$ for any related probability space $(\mathbb{P}, \Omega \times D)$.

Note that $\mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$ always exists.

Definition 7.28 (Conditional Independence)**.** Given a kernel space $(\mathbb{K}, \Omega, D)$, relative probability space $(\mathbb{P}, \Omega \times D)$, variables X,Y and Z, X is *conditionally independent* of Z given Y, written $\mathsf{X} \perp\!\!\!\perp_{\mathbb{K}} \mathsf{Z}|\mathsf{Y}$ if $\mathbb{K}^{\{\mathsf{XY}|\mathsf{Z}\}}$ exists and any of the following equivalent conditions hold:

Almost sure equality

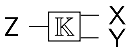
- $\mathbb{P}^{\mathsf{XZ}|\mathsf{Y}} \sim \mathbb{P}^{\mathsf{X}|\mathsf{Y}} \underline{\otimes} \mathbb{P}^{\mathsf{Z}|\mathsf{Y}}$
- For any version of $\mathbb{P}^{\{\mathsf{X}|\mathsf{Y}\}}$, $\mathbb{P}^{\mathsf{X}|\mathsf{Y}} \otimes *_Z$ is a version of $\mathbb{K}^{\{\mathsf{X}|\mathsf{YZ}\}}$
- There exists a version of $\mathbb{K}^{\{\mathsf{X}|\mathsf{YZ}\}}$ constant in Z

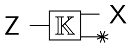
Lemma 7.29 (Diagrammatic consequences of labels)**.** *In general, diagram labels are "well behaved" with regard to the application of any of the special Markov kernels: identities 131, swaps 144, discards 147 and copies 139 as well as with respect to the coherence theorem of the CD category. They are not "well behaved" with respect to composition.*

Fix some Markov kernel space $(\mathbb{K}, \Omega, D)$ *and random variables* X, Y, Z *taking values in* X, Y, Z *respectively.* Sat : *indicates that a labeled diagram satisfies definitions 7.15 and 7.18 with respect to* $(\mathcal{K}, \Omega, D)$ *and* X, Y, Z. *The following always holds:*

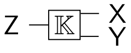
$$\text{Sat} : \mathsf{X} - \mathsf{X} \tag{223}$$

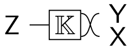
and the following implications hold:

Sat : 

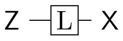
$\Longrightarrow$ Sat : 

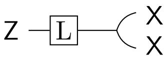
(224)

Sat : 

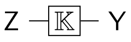
$\Longrightarrow$ Sat : 

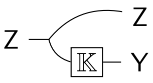
(225)

Sat : 

$\Longrightarrow$ Sat : 

(226)

Sat : 

$\Longrightarrow$ Sat : 

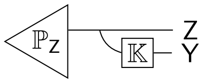
(227)

Proof.
- Id_X is a version of $\mathbb{P}_{\mathsf{X}|\mathsf{X}}$ for all $\mathbb{P}$; $\mathbb{P}_{\mathsf{X}}\mathrm{Id}_X = \mathbb{P}_{\mathsf{X}}$
- $\mathbb{K}\mathrm{Id} \otimes *)(w; A) = \int_{X \times Y} \delta_x(A) \mathbb{1}_Y(y) d\mathbb{K}_w(x, y) = \mathbb{K}_w(A \times Y) = \mathbb{P}_{\mathsf{X}|\mathsf{Z}}(w; A)$

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

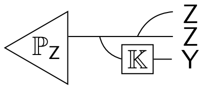
227: Suppose $\mathbb{K}$ is a version of $\mathbb{P}_{\mathsf{Y|Z}}$. Then

$\mathbb{P}_{\mathsf{ZY}} =$



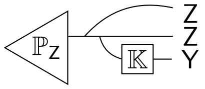
(228)

$\mathbb{P}_{\mathsf{ZZY}} =$



(229)

$=$



(230)

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

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

$$\text{Sat} : \mathsf{Z} -\boxed{\mathbb{K}}- \mathsf{Y}\,,\ \mathsf{Y} -\boxed{\mathbb{L}}- \mathsf{X} \implies \text{Sat} : \mathsf{Z} -\boxed{\mathbb{K}}-\boxed{\mathbb{L}}- \mathsf{X} \tag{231}$$

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

This is an attempt to understand which assumptions lead to see-do models being represented by probability distributions.

8 Decision makers and decision models

We will consider a very general type of decision maker to motivate the construction of decision models. A decision maker takes some kind of evidence - for example, data, prior knowledge or a problem specification - and determines the consequences it believes each decision will have via some *inference rule.* It then examines the consequences of each decision and chooses a preferred decision via some *choice rule.*

More formally, let the set A be a set of possible pieces of evidence a decision maker could obtain, E be the set of consequences it may experience and D the set of decisions it may take. Call any function $C : D \to E$ a *consequence map.* A *inference rule* is a function $f : A \to E^D$ that takes a piece of evidence and returns a consequence map. A *choice rule* $\text{ch} : E^D \to D$ is a function that takes a consequence map and returns a preferred decision. A decision maker is specified by a particular choice of f and ch and implements the *decision function* $h := \text{ch} \circ f$.

Suppose $D = \{0, 1\}$ represents the decision to turn some switch to the on or off positions. Some examples of consequence maps follow:

- $E = \{0, 1\}$ is the state of a light connected to the switch and $\mathbf{C} = \mathrm{Id}_{\{0,1\}}$ is the consequence function that implies the light is always in the same state as the switch
- $E = \{\{0\}, \{1\}, \{0, 1\}\}$ represents sets of possible outcomes and $\mathbf{C} : 0 \mapsto \{0\}, 1 \mapsto \{0, 1\}$ is the consequence function that implies the light is always off when the switch is off, but may take either state with the switch on (i.e. "the bulb may be out")

If $A = \{0, 1\}^{2N}$ is a sequence of joint states of the switch and lightbulb, then an inference function could be the map

$$f(s_1, l_1, .., s_N, l_N) = \begin{cases} \quad 0 \mapsto [\![\frac{\sum_{i\in N} l_i(1-s_i)}{\sum_{i\in N}(1-s_i)} > 0.5]\!] \\ \quad 1 \mapsto [\![\frac{\sum_{i\in N} l_i s_i}{\sum_{i\in N} s_i} > 0.5]\!] \end{cases} \tag{232}$$

That is, f returns the consequence function that maps each switch position to the lightbulb state more commonly observed in conjunction with that switch position in the given data.

Some examples of decision rules, where in each case $\mathbf{C} : D \to E$ is some consequence map:

- $E = \{\text{off}, \text{on}\}$ are states of the light bulb, $u : \text{off} \mapsto 0$, $\text{on} \mapsto 1$ is a utility function and $\text{ch}(\mathbf{C}) = \arg\min_{d\in D} u(\mathbf{C}(d))$
- $E = [0, 1]$ is a set of losses and $\text{ch}(\mathbf{C}) = \arg\min_{d\in D}(\mathbf{C}(d))$
- $E = \{\{0\}, \{1\}, \{0, 1\}\}$ are sets of possible outcomes and ch is the minimax operator: $\text{ch}(\mathbf{C}) = \arg\min_{d\in D} \max_{e\in\mathbf{C}(d)}(e)$

Another example of a decision maker is a learning algorithm that produces binary classifiers. A binary classifier takes some piece of data X and returns a class in $\{0, 1\}$. That is, the set of available decisions D is the set of functions from $X \to \{0, 1\}$. Suppose also that the given information is a set of N labeled examples $A = (X \times \{0, 1\})^N$. Then the inference function f might be an *empirical risk* functional that takes data $a \in A$ and a possible classifier $d \in D$ and returns the number of misclassified examples when d is applied to A. In such a case, ch might be the $\arg\min$ functional, making our decision maker an *empirical risk minimiser.* This is not necessarily a good inference function - typically we are interested in the number of misclassified examples a classifier will produce on unseen data, not the number of misclassified examples it will produce on the data that has already been seen.

8.1 Expected utility maximisers

Expected utility maximisation (henceforth “EU”) necessitates that the class of results models f have a particular signature. The EU choice rule compares a set of probability distributions over states of the world E and returns a decision associated with the preferred probability distribution. Thus, denoting by $\Delta(\mathcal{E})$

the set of probability distribution over E (now equipped with σ-algebra $\mathcal{E}$), f must have the signature $f : A \to \Delta(\mathcal{E})^D$. In particular f returns stochastic functions of the type $D \to \Delta(\mathcal{E})$.

So far we have not made any structural assumptions about the set D. One such assumption that we do want to make is that stochastic decisions are possible. That is, if we can choose $d_1 \in D$ and we can choose $d_2 \in D$, then it is also possible to choose d_1 with probability α and d_2 with probability $(1-\alpha)$ for $0 \leq \alpha \leq 1$. We will make this assumption as follows: D represents an underlying set of "elementary" decisions, and the true set of decisions is the set $\Delta(\mathcal{D})$ of probability measures on D (equipped with some σ-algebra $\mathcal{D}$). Then, for example, if $D = \{d_1, d_2\}$ we can "choose" d_1 by selecting the measure δ_{d_1}, and we can choose a mixture of d_1 and d_2 by selecting the measure $\alpha\delta_{d_1} + (1-\alpha)\delta_{d_2}$. Technically, all that we have done so far is assume that the set of decisions is isomorphic to $\Delta(\mathcal{D})$ - which, if D is standard Borel, ensures that the set of decisions is convex closed. In many cases, we also want the following to hold:

$$f_a(\alpha\delta_{d_1} + (1-\alpha)\delta_{d_2}) = \alpha f_a(\delta_{d_1}) + (1-\alpha) f_a(\delta_{d_2}) \tag{233}$$

That is, we assume that the results model f is *additive*. Informally, choosing to do d_1 with probability α and d_2 with probability $1-\alpha$ will yield the result of d_1 with probability α and the result of d_2 with probability $1-\alpha$. In addition, f_a must be continuous with respect to the total variation norm.

For arbitrary $\gamma \in \Delta(\mathcal{D})$, we can write for all $B \in \mathcal{D}$

$$\gamma(B) = \int_D \delta_d(B) d\gamma(d) \tag{234}$$

Can we also say

$$\gamma = \int_D \delta_d d\gamma(d) \tag{235}$$

For $G \in \mathcal{E}$, write $f_a(\delta_d; C)$ for the probability measure $f_a(\delta_d)$ evaluated at C, and $f_a(\cdot; G)$ for the map $d \mapsto f_a(d; C)$. If $f_a(\cdot; C)$ is measurable for all C and additive, then it is a property of the Lebesgue integral that for all $\gamma \in \Delta(\mathcal{D})$

$$\begin{aligned}\int f_a(\gamma; C) &= \int f_a(\int_D \delta_{d'} d\gamma(d'); C) &(236)\\ &= \int_D f_a(\delta_{d'}; C) d\gamma(d') &(237)\end{aligned}$$

Define $\mathcal{C}_a : D \to \Delta(\mathcal{E})$ by $\mathcal{C}_a : d \mapsto f_a(\delta_d; \cdot)$. Then $\mathcal{C}_a(\cdot; B)$ is measurable and $\mathcal{C}_a(d; \cdot)$ is a probability measure - that is, $\mathcal{C}_a$ is a Markov kernel. We can therefore adopt the product notation defined elsewhere to write the consequences of choosing a distribution γ as $\gamma\mathcal{C}_a$.

Have I made any assumptions here besides additivity, convex closedness of D?

As we are chiefly interested in the set of elementary decisions D', we will henceforth simply use $(D, \mathcal{D})$ for this set.

Assumption 233 still permits some unexpected behaviour with respect to randomised decisions. Consider the following variations of a problem:

Example 8.1 (Fighting children problem)**.** A decision maker is mediating a dispute over a toy between two children. If he chooses to give the toy to the first child, the second child will cry, and if he chooses to give the toy to the second child, the first child will cry (the underlying decision set $D' = \{\text{give to first child}, \text{give to second child}\}$. If he chooses a random procedure:

Version 1 These children strongly object to randomness, and both will cry no matter the outcome

Version 2 These are children with a strong insistence on a version of procedural fairness which no doubt make a great deal of sense to them. In particular, if a procedure is used which gives the first child a chance p of receiving the toy, the first child will cry with probability p irrespective of the outcome. The second child, meanwhile, will cry iff the first child does not

Version 3 Whether or not the children cry depends only on whether or not they receive the toy

In version 1, assumption 233 is violated - the outcome when choosing a random procedure is not simply a mixture of the outcomes of each deterministic decision. In version 3, assumption 233 holds, and the results depend only on the outcome of randomisation, as is desired. In version 2, assumption 233 *also* holds - one child cries iff the other does not, and each child cries with the same probability as the probability that they receive the toy. However, the consequences of randomising and then giving the first child the toy are not the same as simply giving the first child the toy. To discount the possibility of this kind of behaviour, we need additional assumptions.

In the text of this example, we have spoken as if our decision had multiple consequences - whether or not each kid gets a toy, whether or not each kid cries - but the compliance of version 2 with 233 depends on considering only the children crying to be a consequence of our decison. To support the intuition that only version 3 is the kind of problem we want to deal with, we introduce the additional assumption that if we choose an extreme decision δ_d, then one consequence of this decision is that we will have chosen d with probability 1.

Formally, we assume the sample space can be written as $E \times D$ for some E, and, defining the random variable $\mathsf{D} : E \times D \to D$ by the projection $\mathsf{D} : (e, d) \mapsto d$, suppose

$$(\mathcal{C}_{(a,d)})_{\mathsf{D}} = \delta_d \tag{238}$$

We find in Joyce (2000) the assumption of a "supposition function" prob$(\cdot\|\cdot)$ that prob$(C\|C) = 1$ where C is "some distinguished condition". Assumption 238 along with assumption 233 implies an analogous condition.

Theorem 8.2 ([Decision determinism implies decision certainty)**.**

That is, for any stochastic decision γ, the consequence conditional on D is the consequence map f_a itself. This assumption rules out both versions 1 and 2 of the fighting children problem above - it assumes that if we randomise and the result of randomisation is some $d \in D$, this is the same as having chosen d to begin with.

Lemma 8.3 (Conditional agreement implies randomised decisions)**.** *If we have*

$$(\gamma f_a)_{|\mathsf{D}} = f_a \tag{239}$$

Then for all $\gamma \in \Delta(\mathcal{D})$*,*

$$\gamma f_a(G) = \int_D f_a(d'; G) d\gamma(d') \tag{240}$$

Proof. By definition, for $\gamma \in \Delta(\mathcal{D})$ □

A consequence of assumption **??** is that for fixed $a \in A$, every stochastic decision $\gamma \in \Delta(\mathcal{D})$ leads to the same conditional probability $(\gamma f_a)_{|\mathsf{D}}$. In fact, as a result of the assumption that the codomain of f is a set of probability measures as well as assumptions 238 and **??** we find that the tuple $\langle E \times D, \mathcal{E} \otimes \mathcal{D}, \mathcal{D}, f_a \rangle$ for each $a \in A$ forms a *conditional probability space* (Rényi, 1955).

For our causal analysis we work with *subjunctive probability spaces*, a generalisation of the more familiar probability spaces. The subjunctive mood is used to describe hypothetical or supposed states of the world, and subjunctive probability spaces are models that ask for some hypothetical state and give us back a probability space. Subjunctive probability spaces are also different to *conditional probability spaces* Rényi (1956) as *hypothesising* or *supposing* (that is, those things described by the subjunctive mood) are different to *conditioning*, which is more closely analogous to *focussing your attention*.

Does this definition need to be here? I didn't know what this word meant before I looked it up

We use subjunctive probability spaces because *supposition* is a core part of decision problems, one we cannot get away from. In contrast, modelling supposition with conditional probability (which would be necessary if we were to insist on using probability spaces) adds additional structure to our models which isn't clearly warranted and is potentially confusing.

Any decision problem must involve the comparison of different decisions. This comparison takes the form

- *Suppose* I choose the first decision - then the consequence would be X

- *Suppose* I choose the second decision - then the consequence would be Y
- Etc.

If we prefer X to Y, then perhaps we should choose the first decision. We may be ambivalent as to what type of thing X and Y represent, how we ought to determine what values X and Y take or how we ought to determine what is preferable, but it is hard to do away with supposing that different decisions were taken and that these decisions come with their own consequences and still have what could reasonably be considered a decision problem. This process of supposition implicitly invokes a "subjunctive function" - we provide a hypothetical decision, and we are given a consequence of that hypothetical. Of particular interest are models that, for each hypothetical choice, return a probability distribution over space Ω. Such models are called *supposition functions* by Joyce (2000), but we will call them *consequence maps* in this work. We consider this particular type of subjunctive function - from possible decisions to probability distributions - to be attractive because we consider probability to be a sound choice for modelling uncertainty and stochasticity.

Formally, a consequence map $\mathcal{C}$ is a stochastic function or *Markov kernel* from $D \to \Delta(\Omega)$, where $\Delta(\Omega)$ represents the set of all probability distributions on Ω. One might recall that, given two random variables $\mathsf{Y} : \Omega \to Y$ and $\mathsf{X} : \Omega \to X$ on some probability space $(\mathbb{P}, \Omega, \mathcal{F})$, the probability of Y conditional on X is a Markov kernel $X \to \Delta(\mathcal{Y})$. It is possible, in general, to define a probability distribution on the expanded space $\Omega \times D$ such that $\mathcal{C}$ is a conditional probability. However, there are good reasons to keep the concepts of consequence maps and conditional probability separate. Firstly, there are technical issues such as the fact that it is not always possible to find a distribution on $\Omega \times D$ that yields $\mathcal{C}$ as a *unique* conditional probability (Hájek, 2003) (though this requires an uncountable set of possible decisions, which is not a problem we consider in this work). Secondly, it is not clear that that the way we ought to handle consequence maps is identical to the way we handle conditional probabilities. Consider an expanded set of options:

1. Suppose I choose the first decision d_1 - then the consequence would be P_1
2. Suppose I choose the second decision d_2 - then the consequence would be P_2
3. Suppose I choose either the first or second decision d_1 or d_2 - then the consequence would be ???
4. Suppose I choose d_1 with probability $0 \leq q \leq 1$ and d_2 otherwise - then the consequence would be ???

If we regard the consequence map as a conditional probability then it would be natural to consider the result of the third option to be a unique probability distribution obtained by conditioning on $\{d_1, d_2\}$, equal to $\alpha P_1 + (1-\alpha)P_2$ for some fixed $0 \leq \alpha \leq 1$. However, there is no obvious reason that these should be mixed in some fixed proportion α - it seems more appropriate to me to say that if d_1 or d_2 is chosen then the result will be P_1 or P_2.

On the other hand, the result of the fourth option seems like it should be given by $qP_1 + (1-q)P_2$, at least for most ordinary problems. If we were confronted with a mind reader who could tell the difference between us having chosen d_1 and us having randomised between d_1 and d_2 but come up with d_1 anyway then we might have reason to revise this assumption, but we will generally proceed under the assumption that such mind readers are absent. For any ambient probability measure, we can choose q not to be equal to α (as defined in the previous paragraph), and so the result will not be equal to the ambient measure conditioned on $\{d_1, d_2\}$.

In general, choosing the consequence map to be a conditional probability requires there to exist some joint distribution over the decision D and all other random variables in the problem. However, when we adopt hypotheses about what decisions we might make, we throw away key parts of this joint distribution (for example, we throw away the marginal distribution of D) and replace it with whatever we want to suppose instead. Instead of adopting a full joint distribution and then throwing away some parts to get hold of the consequence map, as we would if we were working with a probability space, a subjunctive probability space only supplies those parts which we intend to keep - i.e. in only supplies the consequence map.

9 Notes on category theoretic probability and string diagrams

Category theoretic treatments of probability theory often start with *probability monads* (for a good overview, see (Jacobs, 2018)). A monad on some category C is a functor $T : C \to C$ along with natural transformations called the unit $\eta : 1_C \to T$ and multiplication $\mu : T^2 \to T$. Roughly, functors are maps between categories that preserve identity and composition structure and natural transformations are "maps" between functors that also preserve composition structure. The monad unit is similar to the identity element of a monoid in that application of the identity followed by multiplication yields the identity transformation. The multiplication transformation is also (roughly speaking) associative.

An example of a probability monad is the discrete probability monad given by the functor $\mathcal{D} : \mathbf{Set} \to \mathbf{Set}$ which maps a countable set X to the set of functions from $X \to [0,1]$ that are probability measures on X, denoted $\mathcal{D}(X)$. $\mathcal{D}$ maps a measurable function f to $\mathcal{D}f : X \to \mathcal{D}(X)$ given by $\mathcal{D}f : x \mapsto \delta_{f(x)}$. The unit of this monad is the map $\eta_X : X \to \mathcal{D}(X)$ given by $\eta_X : x \mapsto \delta_x$ (which is equivalent to $\mathcal{D}1_X$) and multiplication is $\mu_X : \mathcal{D}^2(X) \to \mathcal{D}(X)$ where $\mu_X : \mathit{\Omega} \mapsto \sum_\phi \mathit{\Omega}(\phi)\phi$.

For continuous distributions we have the Giry monad on the category **Meas** of mesurable spaces given by the functor $\mathcal{G}$ which maps a measurable space X

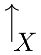
to the set of probability measures on X, denoted $\mathcal{G}(X)$. Other elements of the monad (unit, multiplication and map between morphisms) are the "continuous" version of the above.

Of particular interest is the Kleisli category of the monads above. The Kleisli C_T category of a monad T on category C is the category with the same objects and the morphisms $X \to Y$ in C_T is the set of morphisms $X \to TY$ in C. Thus the morphisms $X \to Y$ in the Kleisli category $\mathbf{Set}_{\mathcal{D}}$ are morphisms $X \to \mathcal{D}(Y)$ in **Set**, i.e. stochastic matrices, and in the Kleisli category $\mathbf{Meas}_{\mathcal{G}}$ we have Markov kernels. Composition of arrows in the Kleisli categories correspond to Matrix products and "kernel products" respectively.

Both $\mathcal{D}$ and $\mathcal{G}$ are known to be *commutative* monads, and the Kleisli category of a commutative monad is a symmetric monoidal category.

Diagrams for symmetric monoidal categories consist of wires with arrows, boxes and a couple of special symbols. The identity object (which we identify with the set $\{*\}$) is drawn as nothing at all $\{*\} := \boxed{\qquad}$ and identity maps are drawn as bare wires:

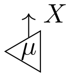
$\mathrm{Id}_X :=$



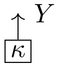
(241)

We draw Kleisli arrows from the unit (i.e. probability distributions) $\mu : \{*\} \to X$ as triangles and Kleisli arrows $\kappa : X \to Y$ (i.e. Markov kernels $X \to \Delta(\mathcal{Y})$) as boxes. We draw the Kleisli arrow $\mathbb{1}_X : X \to \{*\}$ (which is unique for each X) as below

$\mu :=$



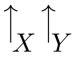
$\kappa :=$



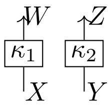
(242)

The product of objects in **Meas** is given by $(X, \mathcal{X}) \cdot (Y, \mathcal{Y}) = (X \times Y, \mathcal{X} \otimes \mathcal{Y})$, which we will often write as just $X \times Y$. Horizontal juxtaposition of wires indicates this product, and horizontal juxtaposition also indicates the tensor product of Kleisli arrows. Let $\kappa_1 : X \to W$ and $\kappa_2 : Y \to Z$:

$(X \times Y, \mathcal{X} \otimes \mathcal{Y}) :=$



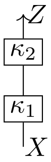
$\kappa_1 \otimes \kappa_2 :=$



(243)

Composition of arrows is achieved by "wiring" boxes together. For $\kappa_1 : X \to Y$ and $\kappa_2 : Y \to Z$ we have

$\kappa_1\kappa_2(x; A) = \int_Y \kappa_2(y; A)\kappa_1(x; dy) :=$



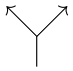
(244)

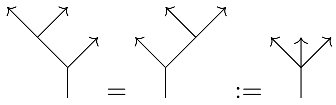
Symmetric monoidal categoris have the following coherence theorem(Selinger, 2010):

Theorem 9.1 (Coherence (symmetric monoidal))**.** *A well-formed equation between morphisms in the language of symmetric monoidal categories follows from the axioms of symmetric monoidal categories ifand only if it holds, up to isomorphism of diagrams, in the graphical language.*

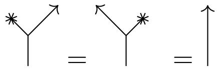
Isomorphism of diagrams for symmetric monoidal categories (somewhat informally) is any planar deformation of a diagram including deformations that cause wires to cross. We consider a diagram for a symmetric monoidal category to be well formed only if all wires point upwards.

In fact the Kleisli categories of the probability monads above have (for each object) unique *copy*: $X \to X \times X$ and *erase*: $X \to \{*\}$ maps that satisfy the *commutative comonoid axioms* that (thanks to the coherence theorem above) can be stated graphically. These differ from the copy and erase maps of *finite product* or *cartesian* categories in that they do not necessarily respect composition of morphisms.

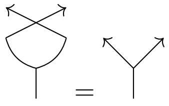
$\text{Erase} = \mathbb{1}_X :=$ $\text{Copy} = x \mapsto \delta_{x,x} :=$  (245)



(246)

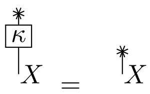


(247)



(248)

Finally, $\{*\}$ is a terminal object in the Kleisli categories of either probability monad. This means that the map $X \to \{*\}$ is unique for all objects X, and as a consequence for all objects X, Y and all $\kappa : X \to Y$ we have



(249)

This is equivalent to requiring for all $x \in X$ $\int_Y \kappa(x;dy) = 1$. In the case of $\mathbf{Set}_{\mathcal{D}}$, this condition is what differentiates a stochastic matrix from a general positive matrix (which live in a larger category than $\mathbf{Set}_{\mathcal{D}}$).

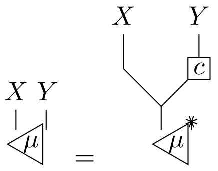
Thus when manipulating diagrams representing Markov kernels in particular (and, importantly, not more general symmetric monoidal categories) diagram isomorphism also includes applications of 246, 247, 248 and 249.

A particular property of the copy map in $\mathbf{Meas}_{\mathcal{G}}$ (and probably $\mathbf{Set}_{\mathcal{D}}$ as well) is that it commutes with Markov kernels iff the markov kernels are deterministic (Fong, 2013).

9.1 Disintegration and Bayesian inversion

Disintegration is a key operation on probability distributions (equivalently arrows $\{*\} \to X$) in the categories under discussion. It corresponds to "finding the conditional probability" (though conditional probability is usually formalised in a slightly different way).

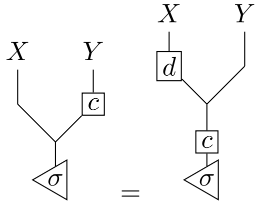
Given a distribution $\mu : \{*\} \to X \otimes Y$, a disintegration $c : X \to Y$ is a Markov kernel that satisfies



(250)

Disintegrations always exist in $\mathbf{Set}_{\mathcal{D}}$ but not in $\mathbf{Meas}_{\mathcal{G}}$. The do exist in the latter if we restrict ourselves to standard measurable spaces. If c_1 and c_2 are disintegrations $X \to Y$ of μ, they are equal μ-A.S. In fact, this equality can be strengthened somewhat - they are equal almost surely with respect to any distribution that shares the "X-marginal" of μ.

Given $\sigma : \{*\} \to X$ and a channel $c : X \to Y$, a Bayesian inversion of (σ, c) is a channel $d : Y \to X$ such that

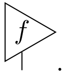


(251)

We can obtain disintegrations from Bayesian inversions and vise-versa.

Clerc et al. (2017) offer an alternative view of Bayesian inversion which they claim doesn't depend on standard measurability conditions, but there is a step in their proof I didn't follow.

9.2 Generalisations

Cho and Jacobs (2019) make use of a larger "CD" category by dropping 249. I'm not completely clear whether you end up with arrows being "Markov kernels for general measures" or something else (can we have negative arrows?). This allows for the introduction of "observables" or "effects" of the form .

Jacobs et al. (2019) make use of an embedding of $\mathbf{Set}_{\mathcal{D}}$ in $\mathbf{Mat}(\mathbb{R}^+)$ with morphisms all positive matrices (I'm not totally clear on the objects, or how they are self-dual - this doesn't seem to be exactly the same as the category of finite dimensional vector spaces). This latter category is compact closed, which - informally speaking - supports the same diagrams as symmetric monoidal categories with the addition of "upside down" wires.

9.3 Key questions for Causal Theories

We will first define *labeled diagrams*. Rather than labelling the wires of our diagrams with *spaces* (as is typical (Selinger, 2010)), we assign a unique label to each "wire segment" (with some qualifications). That is, we assign a unique label to each bare wire in the diagram with the following additonal qualifications:

I'm sure one of the papers I read mentioned labeled diagrams, I just couldn't find it when I looked for it

Since writing this, I found Kissinger (2014) as an example of a diagrammatic system with labeled wires, I will follow it up

- If we have a box in the diagram representing the identity map, the incoming and outgoing wires are given the same label
- If we have a wire crossing in the diagram, the diagonally opposite wires are given the same label
- The input wire and the *two* output wires of the copy map are given the same label

Given two diagrams G_1 and G_2 that are isomorphic under transformations licenced by the axioms of symmetric monoidal categories and commutative comonoid axioms, suppose we have a labelling of G_1. We can label G_2 using the following translation rule:

- For each box in G_2, we can identify a corresponding box in G_1 via labels on each box. For each such pair of boxes, we label the incoming wires of the G_2 box with the labels of the G_1 box preserving the left-right order. We do likewise for outgoing wires.

These rules will lead to a unique labelling of G_2 with all wire segments are labelled. We would like for these rules to yield the following:

- For any sequence of diagram isomorphisms beginning with G_1 and ending with G_2, we end up with the same set of labels
- If we label G_2 according to the rules above then relabel G_1 from G_2 according to the same rules we retrieve the original labels of G_1

We do not prove these properties here, but motivate them via the following considerations:

- These properties obviously hold for the wire segments into and out of boxes
- The only features a diagram may have apart from boxes and wires are wire crossings, copy maps and erase maps
- The labeling rule for wire crossings respects the symmetry of the swap map
- The labeling rule for copy maps respects the symmetry of the copy map and the property described in Equation 248

We will follow the convention whereby "internal" wire labels are omitted from diagrams.

Note also that each wire that terminates in a free end can be associated with a random variable. Suppose for $N \in \mathbb{N}$ we have a kernel $\kappa : A \to \Delta(\times_{i \in N} X_i)$. Define by p_j ($j \in [N]$) the projection map $p_j : \times_{i \in N} X_i \to X_j$ defined by $p_j : (x_0, ..., x_N) \mapsto x_j$. p_j is a measurable function, hence a random variable. Define by π_j the projection kernel $\mathcal{G}(\pi_j)$ (that is, $\pi_j : \mathbf{x} \mapsto \delta_{p_j(\mathbf{x})}$). Note that $\kappa(y; p_j^{-1}(A)) = \int_{X_j} \delta_{p_j(\mathbf{x})}(A) \kappa(y; d\mathbf{x}) = \kappa \pi_j$. Diagrammatically, π_j is the identity map on the j-th wire tensored with the erase map on every other wire. Thus the j-th wire carries the distribution associated with the random variable p_j. We will therefore consider the labels of the "outgoing" wires of a diagram to denote random varaibles (though there are obviously many random variables not represented by such wires). We will additionally distinguish wire labels from spaces by font - wire labels are sans serif $\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{X}, \mathsf{Y}, \mathsf{Z}$ while spaces are serif A, B, C, X, Y, Z.

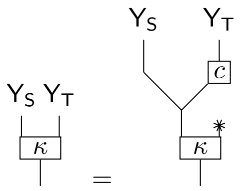
Wire labels appear to have a key advantage over random variables: they allow us to "forget" the sample space as the correct typing is handled automatically by composition and erasure of wires

generalised disintegrations : Of key importance to our work is generalising the notion of disintegration (and possibly Bayesian inversion) to general kernels $X \to Y$ rather than restricting ourselves to probability distributions $\{*\} \to Y$. We will define generalised disintegrations as a straightforward analogy regular disintegrations, but the conditions under which such disintegrations exist are more restrictive than for regular disintegraions.

Definition 9.2 (Label signatures)**.** If a kernel $\kappa : X \to \Delta(Y)$ can be represented by a diagram G with incoming wires $\mathsf{X}_1, ...\mathsf{X}_\mathsf{n}$ and outgoing wires $\mathsf{Y}_1, ..., \mathsf{Y}_\mathsf{m}$, we can assign the kernel a "label signature" $\kappa : \mathsf{X}_1 \otimes ... \otimes \mathsf{X}_\mathsf{n} \dashrightarrow \mathsf{Y}_1 \otimes ... \otimes \mathsf{Y}_m$ or, for short, $\kappa : \mathsf{X}_{[n]} \dashrightarrow \mathsf{Y}_{[m]}$. Note that this signature associates each label with a unique space - the space of X_1 is the space associated with the left-most wire of G and so forth. We will implicitly leverage this correspondence and write with X_1 the space associated with X_1 and so forth. Note that while X_1 is by construction always different from X_2 (or any other label), the space X_1 may coincide with X_2 - the fact that labels always maintain distinctions between wires is the fundamental reason for introducing them in the first place.

There might actually be some sensible way to consider κ to be transforming the measurable functions of a type similar to $\otimes_{i\in[n]}\mathsf{X}_i$ to functions of a type simlar to $\otimes_{i\in[m]}\mathsf{Y}_i$ (or vise versa - perhaps related to Clerc et al. (2017)), but wire labels are all we need at this point

Definition 9.3 (Generalised disintegration)**.** Given a kernel $\kappa : X \to \Delta(Y)$ with label signature $\kappa : \mathsf{X}_{[\mathsf{n}]} \dashrightarrow \mathsf{Y}_{[\mathsf{m}]}$ and disjoint subsets $S, T \subset [m]$ such that $S \cup T = [m]$, a kernel c is a *g-disintigration from S to T* if it's type is compatible with the label signature $c : \mathsf{Y}_\mathsf{S} \dashrightarrow \mathsf{Y}_\mathsf{T}$ and we have the identity (omitting incoming wire labels):



(252)

I have introduced without definition additional labeling operations here: first, each label has a particular space associated with it (in order to license the notion of "type compatible with label signature"), and we have supposed labels can be "bundled".

In contrast to regular disintegrations, generalised disintegrations "usually" do not exist. Consider $X = \{0, 1\}$, $Y = \{0, 1\}^2$ and κ has label signature $\mathsf{X}_1 \dashrightarrow \mathsf{Y}_{\{1,2\}}$ with

$$\kappa : \begin{cases} 1 \mapsto \delta_1 \otimes \delta_1 \\ 0 \mapsto \delta_1 \otimes \delta_0 \end{cases} \tag{253}$$

κ imposes contradictory requirements for any disintegration $c : \{0, 1\} \to \{0, 1\}$ from $\{1\}$ to $\{2\}$: equality for $\mathsf{X}_1 = 1$ requires $c(1; \cdot) = \delta_1$ while equality for $\mathsf{X}_1 = 0$ requires $c(1; \cdot) = \delta_0$. Subject to some regularity conditions (similar to standard Borel conditions for regular disintegrations), we can define g-disintegrations of a canonically related kernel that do generally exist; intuitively, g-disintegrations exist if they take the "input wires" of κ as input wires themselves.

Lemma 9.4. *Given $\kappa : X \to \Delta(Y)$, a kernel $\kappa^\dagger$ is a right inverse iff we have for all $x \in X$, $A \in \mathcal{X}$, $y \in Y$ $\kappa^\dagger(y; A) = \delta_x(A)$, $\kappa(x; \cdot)$-almost surely.*

Proof. Suppose $\kappa^\dagger$ satisfies the almost sure equality for all $x \in X$. Then for all $x \in X$, $A \in \mathcal{X}$ we have $\kappa\kappa^\dagger(x; A) = \int_Y \kappa^\dagger(y; A)\kappa(x; dy) = \int_Y \delta_x(A)\kappa(x; dy) = \delta_x(A)$; that is, $\kappa\kappa^\dagger = \mathrm{Id}_X$, so $\kappa^\dagger$ is a right inverse of κ.

Suppose we have a right inverse $\kappa^\dagger$. By definition, for all $x \in X$ and $A \in \mathcal{X}$ we have $\int_Y \kappa^\dagger(y; A)\kappa(x; dy) = \delta_x(A)$.

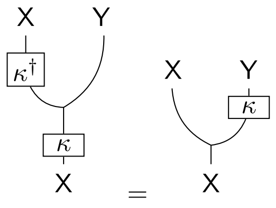
Suppose $x \notin A$ and let $B_\epsilon = \kappa_A^{\dagger-1}((\epsilon, 1])$ for some $\epsilon > 0$. We have $\int_Y \kappa^\dagger(y; A)\kappa(x; dy) = 0 \geq \epsilon\kappa(x; B_\epsilon)$. Thus for any $\epsilon > 0$ we have $\kappa(x; B_\epsilon) = 0$. Consider the set $B_0 = \kappa_A^{\dagger-1}((0, 1])$. For some sequence $\{\epsilon_i\}_{i\in\mathbb{N}}$ such that $\lim_{i\to\infty} \epsilon_i = 0$ we have $B_0 = \cup_{i\in\mathbb{N}} B_{\epsilon_i}$. By countable additivity, $\kappa(x; B_0) = 0$.

Suppose $x \in A$ and let $B^{1-\epsilon} = \kappa_A^{\dagger-1}([0, 1-\epsilon))$. We have $\int_Y \kappa^\dagger(y; A)\kappa(x; dy) = 1 \leq (1-\epsilon)\kappa(x; B^{1-\epsilon}) + 1 - \kappa(x; B^{F.w1-\epsilon}) = 1-\epsilon\kappa(x; B^{1-\epsilon})$. Thus $\kappa(x; B^{1-epsilon}) = 0$ for $\epsilon > 0$. By an argument analogous to the above, we also have $\kappa(x; B^1) = 0$. Thus the $\kappa(x; \cdot)$ measure of the set on which $\kappa^\dagger(y; A)$ disagrees with $\delta_x(A)$ is $\kappa(x; B_0) + \kappa(x; B^1) = 0$ and hence $\kappa^\dagger(y; A) = \delta_x(A)$ $\kappa(x; \cdot)$-almost surely. □

I haven't shown that any map inverting κ implies the existence of a Markov kernel that does so

I am using countable sets below to get my general argument in order without getting too hung up on measurability; I will try to lift it to standard measurable once it's all there

Lemma 9.5. *Given $\kappa : X \to \Delta(Y)$ and a right inverse $\kappa^\dagger$, we have*



(254)

Proof. Let the diagram on the left hand side be L and the diagram on the right hand side be R.

$$L(x; A \times B) = \int_Y \int_{Y\times Y} \mathrm{Id}_Y \otimes \kappa_S^\dagger(y, y'; A \times B)\delta_{(z,z)}(dy \times dy')\kappa\pi_S(x; dz) \tag{255}$$

$$= \int \mathrm{Id}_Y \otimes \kappa^\dagger(z, z; A \times B)\kappa\pi_S(x; dz) \tag{256}$$

$$= \int \delta_z(A)\kappa_S^\dagger(z; B)\kappa\pi_S(x; dz) \tag{257}$$

$$= \int_A \kappa_S^\dagger(z; B)\kappa\pi_S(x; dz) \tag{258}$$

$$= \delta_x(B)\kappa\pi_S(x; A) \tag{259}$$

Where 259 follows from Lemma 9.4.

$$R(x; A \times B) = \int \delta_{(x,x)}(dy \times dy')\kappa\pi_S \otimes \mathrm{Id}_X(y, y'; A \times B) \tag{260}$$

$$= \kappa\pi_S(x; A)\delta_x(B) \qquad = L \tag{261}$$

□

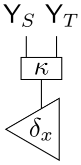
Theorem 9.6. *Given countable X and standard measurable Y, $n, m \in \mathbb{N}$, $S, T \subset [m]$, κ with label signature $\mathsf{X}_{[\mathsf{n}]} \dashrightarrow \mathsf{Y}_{[\mathsf{m}]}$ a g-disintegration exists from S to T if $\kappa\pi_S$ is right-invertible*

via a Markov kernel

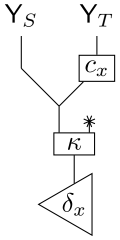
Proof. In addition, as R is a composition of Markov kernels, and hence a Markov kernel itself, L must also be a Markov kernel even if $\kappa^\dagger$ is not.

For all $x \in X$ we have a (regular) disintegration $c_x : Y_S \to \Delta(Y_T)$ of $\kappa(x; \cdot)$ by standard measurability of Y. Define $c : X \otimes Y_S \to \Delta(Y_T)$ by $c : (x, y_S) \mapsto c_x(y_S)$. Clearly, $c(x, y_S)$ is a probability distribution on Y_T for all $(x, y_S) \in X \otimes Y_S$. It remains to show $c(\cdot)^{-1}(B)$ is measurable for all $B \in \mathcal{B}([0, 1])$. But $c(\cdot)^{-1}(B) = \cap_{x \in X} c_y(\cdot)^{-1}(B)$. The right hand side is measurable by measurability of $c_y(\cdot)^{-1}(B)$ countability of X, so c is a Markov kernel.

By the definition of c_x, we have for all $x \in X$

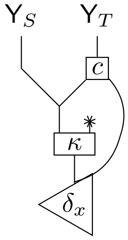


$=$



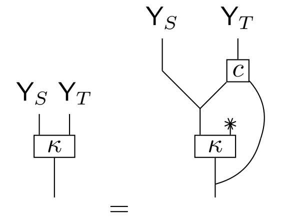
(262)

$=$



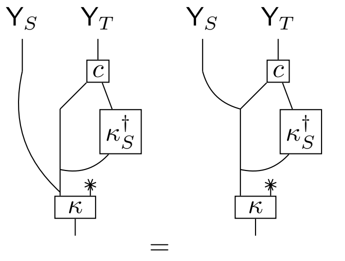
(263)

Which implies

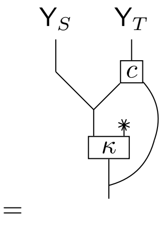


(264)

Finally, we have



(265)



(266)

Where the first line follows from 247 and the second line from 254. If $\kappa_S^\dagger$ is a Markov kernel, then $\curlyvee(\mathrm{Id}_{Y_S} \otimes \kappa_S^\dagger)c$ is a g-disintegration. □

In the reverse direction, suppose κ is such that $\kappa\pi_T = \mathrm{Id}_X$; that is, π_T is a right inverse of κ. If $\kappa\pi_S$ is not right invertible then, by definition, there is no d such that $\kappa\pi_S d\pi_T = \mathrm{Id}_X$. However, if a g-disintegration of κ exists then there is a d such that $\kappa\pi_S d = \kappa$, a contradiction. Thus if $\kappa\pi_S$ is not right invertible then there is *in general* no g-disintegration from S to T.

References

Joshua D. Angrist and Jörn-Steffen Pischke. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press, Princeton ; Oxford, with french flaps edition edition, December 2014. ISBN 978-0-691-15284-4.

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`.

A. Philip Dawid. Decision-theoretic foundations for statistical causality. *arXiv:2004.12493 [math, stat]*, April 2020. URL `http://arxiv.org/abs/2004.12493`. arXiv: 2004.12493.

Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2–21, August 2018. ISSN 0277-9536. doi: 10.1016/j.socscimed.2017.12.005. URL `http://www.sciencedirect.com/science/article/pii/S0277953617307359`.

Ronald A. Fisher. Cancer and Smoking. *Nature*, 182(4635):596–596, August 1958. ISSN 1476-4687. doi: 10.1038/182596a0. URL `https://www.nature.com/articles/182596a0`. Number: 4635 Publisher: Nature Publishing Group.

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.

D. Heckerman and R. Shachter. Decision-Theoretic Foundations for Causal Reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, December 1995. ISSN 1076-9757. doi: 10.1613/jair.202. URL `https://www.jair.org/index.php/jair/article/view/10151`.

James J. Heckman. Randomization and Social Policy Evaluation. SSRN Scholarly Paper ID 995151, Social Science Research Network, Rochester, NY, July 1991. URL `https://papers.ssrn.com/abstract=995151`.

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.

Alan Hájek. Interpretations of Probability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019. URL https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/.

Bart Jacobs. From probability monads to commutative effectuses. *Journal of Logical and Algebraic Methods in Programming*, 94:200–237, January 2018. ISSN 2352-2208. doi: 10.1016/j.jlamp.2016.11.006. URL http://www.sciencedirect.com/science/article/pii/S2352220816301122.

Bart Jacobs, Aleks Kissinger, and Fabio Zanasi. Causal Inference by String Diagram Surgery. In Mikoaj Bojaczyk and Alex Simpson, editors, *Foundations of Software Science and Computation Structures*, Lecture Notes in Computer Science, pages 313–329. Springer International Publishing, 2019. ISBN 978-3-030-17127-8.

James M. Joyce. Why We Still Need the Logic of Decision. *Philosophy of Science*, 67:S1–S13, 2000. ISSN 0031-8248. URL www.jstor.org/stable/188653.

Aleks Kissinger. Abstract Tensor Systems as Monoidal Categories. In Claudia Casadio, Bob Coecke, Michael Moortgat, and Philip Scott, editors, *Categories and Types in Logic, Language, and Physics: Essays Dedicated to Jim Lambek on the Occasion of His 90th Birthday*, Lecture Notes in Computer Science, pages 235–252. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. ISBN 978-3-642-54789-8. doi: 10.1007/978-3-642-54789-8_13. URL https://doi.org/10.1007/978-3-642-54789-8_13.

Chayakrit Krittanawong, Bharat Narasimhan, Zhen Wang, Joshua Hahn, Hafeez Ul Hassan Virk, Ann M. Farrell, HongJu Zhang, and WH Wilson Tang. Association between chocolate consumption and risk of coronary artery disease: a systematic review and meta-analysis:. *European Journal of Preventive Cardiology*, July 2020. doi: 10.1177/2047487320936787. URL http://journals.sagepub.com/doi/10.1177/2047487320936787. Publisher: SAGE PublicationsSage UK: London, England.

Finnian Lattimore and David Rohde. Replacing the do-calculus with Bayes rule. *arXiv:1906.07125 [cs, stat]*, December 2019. URL http://arxiv.org/abs/1906.07125. arXiv: 1906.07125.

David K Lewis. Causation. *Journal of Philosophy*, 1986.

Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, New York, NY, 2 edition edition, November 2014. ISBN 978-1-107-69416-3.

Naomi Oreskes and Erik M. Conway. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Climate Change: How a Handful of Scientists ... Issues from Tobacco Smoke to Global Warming*. Bloomsbury Press, New York, NY, June 2011. ISBN 978-1-60819-394-3.

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

Judea Pearl. Challenging the hegemony of randomized controlled trials: A commentary on Deaton and Cartwright. *Social Science & Medicine*, 2018a.

Judea Pearl. Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes. *Journal of Causal Inference*, 6(2), 2018b. 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`.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 1 edition edition, May 2018. ISBN 978-0-465-09760-9.

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

Robert N. Proctor. The history of the discovery of the cigarettelung cancer link: evidentiary traditions, corporate denial, global toll. *Tobacco Control*, 21(2):87–91, March 2012. ISSN 0964-4563, 1468-3318. doi: 10.1136/tobaccocontrol-2011-050338. URL `https://tobaccocontrol.bmj.com/content/21/2/87`. Publisher: BMJ Publishing Group Ltd Section: The shameful past.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

Donald B. Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, March 2005. ISSN 0162-1459. doi: 10.1198/016214504000001880. URL `https://doi.org/10.1198/016214504000001880`.

A. Rényi. On Conditional Probability Spaces Generated by a Dimensionally Ordered Set of Measures. *Theory of Probability & Its Applications*, 1(1):55–64, January 1956. ISSN 0040-585X. doi: 10.1137/1101005. URL `https://epubs.siam.org/doi/abs/10.1137/1101005`.

Alfréd Rényi. On a new axiomatic theory of probability. *Acta Mathematica Academiae Scientiarum Hungarica*, 6(3):285–335, September 1955. ISSN 1588-2632. doi: 10.1007/BF02024393. URL `https://doi.org/10.1007/BF02024393`.

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.

Ilya Shpitser and Judea Pearl. Complete Identification Methods for the Causal Hierarchy. *Journal of Machine Learning Research*, 9(Sep):1941–1979, 2008. ISSN ISSN 1533-7928. URL https://www.jmlr.org/papers/v9/shpitser08a.html.

Statista. Cigarettes - worldwide | Statista Market Forecast, 2020. URL https://www.statista.com/outlook/50010000/100/cigarettes/worldwide.

Robert Wiblin. Why smoking in the developing world is an enormous problem and how you can help save lives, 2016. URL https://80000hours.org/problem-profiles/tobacco/.

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/.

World Health Organisation. Tobacco Fact sheet no 339, 2018. URL https://www.webcitation.org/6gUXrCDKA.

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: