When does one variable have a probabilistic causal effect on another?

David Johnston

November 5, 2021

Contents

1 Introduction

Two widely used approaches to causal modelling are *graphical causal models* and *potential outcomes models*. Graphical causal models, which include Causal Bayesian Networks and Structural Causal Models, provide a set of *intervention* operations that take probability distributions and a graph and return a modified probability distribution (Pearl, 2009). Potential outcomes models feature *potential outcome variables* that represent the "potential" value that a quantity of interest would take under the right circumstances, a potential that may be realised if the circumstances actually arise, but will otherwise remain only a potential or *counterfactual* value (Rubin, 2005).

One challenge for both of these approaches is understanding how their causal primitives – interventions and potential outcome variables respectively – relate to the causal questions we are interested in. This challenge is related to the distinction, first drawn by (Korzybski, 1933), between "the map" and "the territory". Causal models, like other models, are "maps" that purport to represent a "territory" that we are interested in understanding. Causal primitives are elements of the maps, and the things to which they refer are parts of the territory. The maps contain all the things that we can talk about unambiguously, so it is challenging to speak clearly about how parts of the maps relate to parts of the territory that fall outside of the maps.

For example, Hernán and Taubman (2008), who observed that many epidemiological papers have been published estimating the "causal effect" of body mass index and argued that, because *actions* affecting body mass index[1] are vaguely defined, potential outcome variables and causal effects themselves become ill-defined. We note that "actions targeting body mass index" are not elements of a potential outcomes model but "things to which potential outcomes should correspond". The authors claim is that vagueness in the "territory" leads to ambiguity about elements of the "map" – and, as we have suggested, anything we can try to say about the territory is unavoidably vague. This seems like a serious problem.

In a response, Pearl (2018) argues that *interventions* (by which we mean the operation defined by a causal graphical model) are well defined, but may not always be a good model of an action. Pearl further suggests that interventions in graphical models correspond to "virtual interventions" or "ideal, atomic interventions", and that perhaps carefully chosen interventions can be good models of actions. Shahar (2009), also in response, argued that interventions targeting body mass index applied to correctly specified graphical causal models will necessarily yield no effect on anything else which, together with Pearl's suggestion, implies perhaps that an "ideal, atomic intervention" on body mass index cannot have any effect on anything else. If this is so, it seems that we are dealing with

[1] the authors use the term "intervention", but they do not use it mean a formal operation on a graphical causal model, and we reserve the term for such operations to reduce ambiguity.

quite a serious case of vagueness – there is a whole body of literature devoted to estimating a "causal effect" that, it is claimed, is necessarily equal to zero! Authors of the original literature on the effects of BMI might counter that they were estimating something different that wasn't necessarily zero, but as far as we are concerned such a response would only underscore the problem of ambiguity.

One of the key problems in this whole discussion is how the things we have called *interventions* – which are elements of causal models – relate to the things we have called *actions*, which live outside of causal models. One way to address this difficulty is to construct a bigger causal model that can contain both "interventions" and "actions", and we can then speak unambiguously about how one relates to another. This is precisely what we do here.

- We need to talk about variables
- We use compatibility + string diagrams
- We consider causation in terms of "proxy control"

2 Variables and Probability Models

2.1 Semantics of observed and unobserved variables

We are interested in constructing *probabilistic models* which explain some part of the world. In a model, variables play the role of "pointing to the parts of the world the model is explaining". Both observed an unobserved variables play important roles in causal modelling and we think it is worth clarifying what variables of either type refer to. Ultimately, our interpretation is largely the standard one: a probabilistic model is associated with an experiment or measurement procedure that yields values in a well-defined set. Observable variables are obtained by applying well-defined functions to the result of this total measurement. We explain how we can use a richer sample space that includes unobserved variables. Unobserved variables are formally modelled in exactly the same way as observed variables, but unlike observed variables we don't offer a standard interpretation of unobserved variables.

Consider Newton's second law in the form $\mathcal{F} = \mathcal{M}\mathcal{A}$ as a simple example of a model that relates variables $\mathcal{F}$, $\mathcal{M}$ and $\mathcal{A}$. As Feynman (1979) noted, this law is incomplete – in order to understand it, we must bring some pre-existing understanding of force, mass and acceleration as independent things. Furthermore, the nature of this knowledge is somewhat perculiar. Acknowledging that physicists happen to know a great deal about determining the forces on an object, it remains true that in order to actually say what the net force on a real object is, even a highly knowledgeable physicist will still have to go and do some measurements, and the result of such measurements will be a vector representing the net forces on that object.

This suggests that we can think about "force" $\mathcal{F}$ (or mass or acceleration) as a kind of procedure that we apply to a particular real world object and which

returns a mathematical object (in this case, a vector). We will call $\mathcal{F}$ a *procedure*. Our view of $\mathcal{F}$ is akin to Menger (2003)'s notion of variables as "consistent classes of quantities" that consist of pairing between real-world objects and quantities of some type. Force $\mathcal{F}$ itself is not a well-defined mathematical thing, as measurement procedures are not mathematically well-defined. At the same time, the set of values it may yield *are* well-defined mathematical things.

We will assume that any procedure will eventually yield an unambiguous value in a defined mathematical set. No actual procedure can be guaranteed to have this property – any apparatus, however robust, could suffer catastrophic failure – but we assume that we can study procedures reliable enough that we don't lose much by making this assumption. This assumption allows us to say a procedure $\mathcal{B}$ yields values in B. $\mathcal{B} \bowtie x$ is the proposition that $\mathcal{B}$, when completed, yields the value $x \in B$, and by assumption exactly one of these propositions is true. For $A \subset B$, $\mathcal{B} \bowtie A$ is the proposition ${}_{x \in A}\mathcal{B} \bowtie x$. Two procedures $\mathcal{B}$ and $\mathcal{C}$ are the same if $\mathcal{B} \bowtie x \iff \mathcal{C} \bowtie x$ for all $x \in B$.

The notion of "yielding values" allows us to define an operation akin to function composition. If I have a procedure $\mathcal{B}$ that takes values in some set B, and a function $f : B \to C$, define the "composition" $f \circ \mathcal{B}$ to be the procedure $\mathcal{C}$ that yields $f(x)$ whenever $\mathcal{B}$ yields x. For example, $\mathcal{MA}$ is the composition of $h : (x, y) \mapsto xy$ with the procedure $(\mathcal{M}, \mathcal{A})$ that yields the mass and acceleration of the same object. Composition is associative - for all $x \in B$:

$$
\begin{aligned}
(g \circ f) \circ \mathcal{B} \text{ yields } x &\iff B \text{ yields } (g \circ f)^{-1}(x) && (1)\\
&\iff B \text{ yields } f^{-1}(g^{-1}(x)) && (2)\\
&\iff f \circ B \text{ yields } g^{-1}(x) && (3)\\
&\iff g \circ (f \circ B) \text{ yields } x && (4)
\end{aligned}
$$

One might whether there is also some kind of "append" operation that takes a standalone $\mathcal{M}$ and a standalone $\mathcal{A}$ and returns a procedure $(\mathcal{M}, \mathcal{A})$. Unlike function composition, this would be an operation that acts on two procedures rather than a procedure and a function. Rather than attempt to define any operation of this type, we simply assume that somehow a procedure has been devised that measures everything of interest, which we will call $\mathcal{S}$ which takes values in Ψ. We assume $\mathcal{S}$ is such that any procedure of interest can be written as $f \circ \mathcal{S}$ for some f.

For the model $\mathcal{F} = \mathcal{MA}$, for example, we could assume $\mathcal{F} = f \circ \mathcal{S}$ for some f and $(\mathcal{M}, \mathcal{A}) = g \circ \mathcal{S}$ for some g. In this case, we can get $\mathcal{MA} = h \circ (\mathcal{M}, \mathcal{A}) = (h \circ g) \circ \mathcal{S}$. Note that each procedure is associated with a unique function with domain Ψ.

Thus far, Ψ is a "sample space" that only contains observable variables. To include unobserved variables, we posit a richer sample space Ω such that the measurement $\mathcal{S}$ determines an element of some partition of Ω rather than an element of Ω itself. Then, by analogy to procedures defined with respect to $\mathcal{S}$, we identify variables in general with measurable functions defined on the domain Ω.

Specifically, suppose $\mathcal{S}$ takes values in Ψ. Then we can propose a sample space Ω such that $|\Omega| \geq |\Psi|$ and a surjective function $\mathsf{S} : \Omega \to \Psi$ associated with $\mathcal{S}$. We connect Ω, S and $\mathcal{S}$ with the notion of *consistency with obseravation*:

$\omega \in \Omega$ is *consistent with observation* iff the result yielded by $\mathcal{S}$ is equal to $\mathsf{S}(\omega)$ (5)

Thus the procedure $\mathcal{S}$ eventually restricts the observationally consistent elements of Ω. If $\mathcal{S}$ yield the result s, then the consistent values of Ω will be $\mathsf{S}^{-1}(s)$.

One thing to note in this setup is that two different sets of measurement outcomes Ψ and Ψ' entail a different mesurement procedures $\mathcal{S}$ and $\mathcal{S}'$, but different sample spaces Ω and Ω' may be used to model a single procedure $\mathcal{S}$. We will sometimes consider different models of the same observable procedures.

As far as we know, distinguishing variables from procedures is somewhat nonstandard, but it is a useful distinction to make. While they may not be explicitly distinguished, both variables and procedures are often discussed in statistical texts. For example, Pearl (2009) offers the following two, purportedly equivalent, definitions of variables:

> By a *variable* we will mean an attribute, measurement or inquiry that may take on one of several possible outcomes, or values, from a specified domain. If we have beliefs (i.e., probabilities) attached to the possible values that a variable may attain, we will call that variable a random variable.
>
> This is a minor generalization of the textbook definition, according to which a random variable is a mapping from the sample space (e.g., the set of elementary events) to the real line. In our definition, the mapping is from the sample space to any set of objects called "values," which may or may not be ordered.

Our view is that the first definition is a definition of a procedure, while the second is a definition of a variable. Variables model procedures, but they are not the same thing. We can establish this by noting that, under our definition, every procedure of interest – that is, all procedures that can be written $f \circ \mathcal{S}$ for some f – is modeled by a variable, but there may be variables defined on Ω that do not factorise through $\mathcal{S}$, and these variables do not model procedures.

2.2 Events

To recap, we have a procedure $\mathcal{S}$ yielding values in Ψ that measures everything we are interested in, a sample space Ω and a function S that models $\mathcal{S}$ in the sense of Definition 5. We assume also that Ψ has a σ-algebra $\mathcal{E}$ (this may be the power set of Ψ, as measurement procedures are typically limited to finite precision). Ω is equipped with a σ-algebra $\mathcal{F}$ such that $\sigma(\mathsf{S}) \subset \mathcal{F}$. If a procedure $\mathcal{X} = f \circ \mathcal{S}$ then we define $\mathsf{X} : \Omega \to X$ by $\mathsf{X} := f \circ \mathsf{S}$.

If a particular procedure $\mathcal{X} = f \circ \mathcal{S}$ eventually yields a value x, then the values of Ω consistent with observation must be a subset of $\mathsf{X}^{-1}(x)$. We define an *event* $\mathsf{X} \bowtie x :\equiv \mathsf{X}^{-1}(x)$, which we read "the event that X yields x". An event $\mathsf{X} \bowtie x$ occurs if the consistent values of Ω are a subset of $\mathsf{X} \bowtie x$, thus "the event that X yields x occurs$\equiv$ $\mathcal{X}$ yields x". The definition of events applies to all types of variables, not just observables, but we only provide an interpretation of events "occurring" when the variable X is associated with some $\mathcal{X}$.

For measurable $A \in \mathcal{X}$, $\mathsf{X} \bowtie A = \bigcup_{x \in A} \mathsf{X} \bowtie x$.

Given $\mathsf{Y} : \Omega \to X$, we can define an append operation for variables: $(\mathsf{X}, \mathsf{Y}) := \omega \mapsto (\mathsf{X}(\omega), \mathsf{Y}(\omega))$. (X, Y) has the property that $(\mathsf{X}, \mathsf{Y}) \bowtie (x, y) = \mathsf{X} \bowtie x \cap \mathsf{Y} \bowtie y$, which supports the interpretation of (X, Y) as the values yielded by X and Y together.

It is common to use the symbol "=" instead of "$\bowtie$", but we want to avoid this because $\mathsf{Y} = y$ already has a meaning, namely that Y is a constant function everywhere equal to y.

2.3 Probabilistic models for causal inference

We have a "skeletal model" process $\mathcal{S}$, a sample space $(\Omega, \mathcal{F})$ and a collection of *variables*, which we write with the sans serif font: X, Y_i. The process $\mathcal{S}$ is related to the model Ω and the observable variables via the notion of *consistency with observation*. To this skeleton, we want to add a model that relates variables probabilistically. Without delving into it too deeply, such a model can be thought to be something like a quantitative forecast of the results eventually yielded by $\mathcal{S}$.

For causal inference, we need to modify the standard approach to constructing probability models on a sample space $(\Omega, \mathcal{F})$ to allow for *holes* in our model. Hájek (2003) defines *probability gaps* as propositions that do not have a probability assigned to them (these propositions might correspond to events in the sense discussed above), and probability holes are a particular kind of probability gap. Before defining models with holes in them, we will consider the example of *truncated factorisation* to illustrate the need for models of this type.

For this example, we will assume that the reader is familiar enough with marginal probabilities, conditional probabilities and causal models to follow along. We will offer more careful definitions of terms subsequently.

Suppose we have a causal Bayesian network $(\mathbb{P}^{\mathsf{XYZ}}, \mathcal{G})$ where $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and $\mathsf{Z} : \Omega \to Z$ are variables, $\mathbb{P}^{\mathsf{XYZ}}$ is a probability measure on $X \times Y \times Z$ that we call "a probability model of $(\mathsf{X}, \mathsf{Y}, \mathsf{Z})$" and $\mathcal{G}$ is a Directed Acyclic Graph whose vertices we identify with X, Y and Z that contains the edges $\mathsf{X} \longrightarrow \mathsf{Y}$ and $\mathsf{X} \longleftarrow \mathsf{Z} \longrightarrow \mathsf{Y}$. "Setting X to x" is an operation that takes as inputs $\mathbb{P}^{\mathsf{XYZ}}$, $\mathcal{G}$ and some $x \in X$ and returns a new probability measure $\mathbb{P}^{\mathsf{XYZ}}_x$ on $X \times Y \times Z$ given by (Pearl, 2009, page 24):

$$\mathbb{P}^{\mathsf{XYZ}}_x(x', y, z) = \mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}(y|x, z)\mathbb{P}^{\mathsf{Z}}(z)[\![x = x']\!] \tag{6}$$

Equation 6 embodies three assumptions about a model of the operation of "setting X to x". First, such a model must assign probability 1 to the proposition

that X yields x. Second, such a model must assign the same marginal probability distribution to Z as the input distribution; $\mathbb{P}^{\mathsf{Z}} = \mathbb{P}^{\mathsf{Z}}_x$. Finally, our model must also assign the same conditional probability to Y given X and Z; $\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}} = \mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}_x$.

Notice that the map $x \mapsto \mathbb{P}^{\mathsf{XYZ}}_x$ itself has the type of a conditional probability - that is, it takes some $x \in X$ and returns a probability distribution over $(\mathsf{X}, \mathsf{Y}, \mathsf{Z})$. In fact, a popular notation for this map suggests that it is a conditional probability of sorts: $\mathbb{P}^{\mathsf{XYZ}|do(\mathsf{X}=x)}$. Suppose that this actually is a conditional probability. That is, suppose X is an observable associated with $\mathcal{X}$ and I have some choice available that I believe can dictate the value yielded by $\mathcal{X}$ (alternatively, I can dictate that $\mathsf{X} \bowtie x$ occurs for any $x \in X$), and furthermore I have some $\mathcal{U}$ (modeled by U), taking values in X, that yields whatever choice I end up making. Then I can define a probabilisitic model $\mathbb{Q}^{\mathsf{XYZ}|\mathsf{U}} := x \mapsto \mathbb{P}^{\mathsf{XYZ}}_x$.

Now, the result yielded by $\mathcal{U}$ is a matter of choice, and I may be perfectly capable of making this choice without specifying a probability distribution $\mathbb{Q}^{\mathsf{U}}$. If so, I can leave $\mathbb{Q}^{\mathsf{U}}$ as a *probability hole*, and $\mathbb{Q}^{\mathsf{XYZ}|\mathsf{U}}$ is thus a *probability model with a hole*.

There are a few features of $\mathbb{Q}^{\mathsf{XYZ}|\mathsf{U}}$ we want to point out. Firstly, $\mathbb{Q}^{\mathsf{XYZ}|\mathsf{U}}$ should be unique, but Equation 6 does not necessarily yield a unique $\mathbb{P}^{\mathsf{XYZ}}_x$ for each x. If there are measure zero sets in the range of (X, Z) then $\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}$ can be chosen arbitrarily on these sets, and because these do not necessarily correspond to zero measure sets in the range of Z, different choices can lead to different versions of $\mathbb{P}^{\mathsf{XYZ}}_x$.

Secondly, $\mathbb{Q}^{\mathsf{XYZ}|\mathsf{U}}$ should be valid, but such that Equation 6 does not always yield a valid probability model. For example, if $\mathsf{X} = \mathsf{Z}$ – which is to say, X and Z are the same function on Ω – then any probability model must assign probability 1 to the event $(\mathsf{X}, \mathsf{Z}) \bowtie \{(x, x) | x \in X\}$. However, this cannot be done in accordance with Equation 6 for all $x \in X$ if $|X| \geq 2$ because for $x \neq x' \in X$, $\mathbb{P}^{\mathsf{X}}_x \neq \mathbb{P}^{\mathsf{X}}_{x'}$ while $\mathbb{P}^{\mathsf{Z}}_x = \mathbb{P}^{\mathsf{Z}}_{x'}$.

Finally, $\mathbb{Q}^{\mathsf{XYZ}|\mathsf{U}}$ should be extendable to a valid $\mathbb{Q}^{\mathsf{XYZU}}_\alpha$ for every $\mathbb{Q}^{\mathsf{U}}_\alpha$ that "fills the hole".

To construct probability models with holes in them, we need to address these three issues – uniqueness, validity and valid extendability.

The is an extension of the point (2), I think it's interesting but also a complication. Not quite sure where to put it

The example of $\mathsf{X} = \mathsf{A}$ might seem absurd. Consider instead $\mathsf{Z} = (\mathsf{H}, \mathsf{W})$, representing the height in metres and weight in kilograms of a particular person, and X represents their body mass index, which is to say $\mathsf{X} = \frac{\mathsf{W}}{\mathsf{H}^2}$ (not that in both cases we are using "=", which means these variables are *equal as functions*, not that they *yield the same result with probability 1*, which would involve the symbol "$\bowtie$"). A causal graph with exactly these variables and arrows analogous to ours appears in Shahar (2009). However, generally there is no $\mathbb{P}^{\mathsf{XHW}}_x$ that satisfies both $\mathsf{X} \bowtie \frac{\mathsf{W}}{\mathsf{H}^2}$ with probability 1 and Equation 6. This is true, for example, whenever $\mathbb{P}^{\mathsf{X}}$ has support at more than one point.

2.4 Markov kernels and string diagrams

We say, given a variable $\mathsf{X} : \Omega \to X$, a probability distribution $\mathbb{P}^{\mathsf{X}}$ is a probability measure on $(X, \mathcal{X})$. Recall that a probability measure is a σ-additive function $\mathbb{P}^{\mathsf{X}} : \mathcal{X} \to [0, 1]$ such that $\mathbb{P}^{\mathsf{X}}(\emptyset) = 0$ and $\mathbb{P}^{\mathsf{X}}(X) = 1$. Given a second variable $\mathsf{Y} : \Omega \to Y$, a conditional probability $\mathbb{Q}^{\mathsf{X}|\mathsf{Y}}$ is a Markov kernel $\mathbb{Q}^{\mathsf{X}|\mathsf{Y}} : X \twoheadrightarrow Y$which is a map $Y \times \mathcal{X} \to [0, 1]$ such that

1. $y \mapsto \mathbb{Q}^{\mathsf{X}|\mathsf{Y}}(A|y)$ is $\mathcal{B}$-measurable for all $A \in \mathcal{X}$
2. $A \mapsto \mathbb{Q}^{\mathsf{X}|\mathsf{Y}}K(A|y)$ is a probability measure on $(X, \mathcal{X})$ for all $y \in Y$

If we consider only variables takeing values in discrete sets, Markov kernels and probability distributions have simple representations: a Markov kernel is a positive Matrix with row sum 1 and a probability distribution is a positive row vector with sum 1. Like matrices, we can define products between Markov kernels and in the case of discrete sets, the product of two Markov kernels is precisely the matrix product.

To help us work with Markov kernels, we introduce the string diagram notation taken from the study of Markov categories. Markov categories are abstract categories that represent models of the flow of information. We can form Markov categories from collections of sets – for example, discrete sets or standard measurable sets – along with the Markov kernel product as the composition operation. Markov categories come equipped with a graphical language of *string diagrams*, and a coherence theorem which states that calid proofs using string diagrams correspond to valid theorems in *any* Markov category (Selinger, 2010). More comprehensive introductions to Markov categories can be found in Fritz (2020); Cho and Jacobs (2019).

We use the graphical language to assist in expressing the concept of probability models with holes.

In the graphical language, Markov kernels are drawn as boxes with input and output wires, and probability measures (which are kernels with the domain $\{*\}$) are represented by triangles:

$$\mathbf{K} := \quad -\boxed{\mathbf{K}}- \tag{7}$$

$$\mathbf{P} := \triangleleft\!\boxed{\mathbf{P}}- \tag{8}$$

Two Markov kernels $\mathbf{L} : X \twoheadrightarrow Y$ and $\mathbf{M} : Y \twoheadrightarrow Z$ have a product $\mathbf{LM} : X \twoheadrightarrow Z$, given in the discrete case by the matrix product $\mathbf{LM}(z|x) = \sum_{y \in Y} \mathbf{M}(z|y)\mathbf{L}(y|x)$. Graphically, we represent products between compatible Markov kernels by joining wires together:

$$\mathbf{LM} := \; X \; \boxed{\mathbf{K}}\!-\!\boxed{\mathbf{M}} \; Z \tag{9}$$

The Cartesian product $X \times Y := \{(x, y) | x \in X, y \in Y\}$. Given kernels $\mathbf{K} : W \twoheadrightarrow Y$ and $\mathbf{L} : X \twoheadrightarrow Z$, the tensor product $\mathbf{K} \otimes \mathbf{L} : W \times X \twoheadrightarrow Y \times Z$

given by $(\mathbf{K} \otimes \mathbf{L})(y, z|w, x) := K(y|w)L(z|x)$. The tensor product is graphically represeted by drawing kernels in parallel:

$$\mathbf{K} \otimes \mathbf{L} := \begin{array}{l} W \text{-}\boxed{\mathbf{K}}\text{-} Y \\ X \text{-}\boxed{\mathbf{L}}\text{-} Z \end{array} \tag{10}$$

We read diagrams from left to right (this is somewhat different to Fritz (2020); Cho and Jacobs (2019); Fong (2013) but in line with Selinger (2010)), and any diagram describes a set of nested products and tensor products of Markov kernels. There are a collection of special Markov kernels for which we can replace the generic "box" of a Markov kernel with a diagrammatic elements that are visually suggestive of what these kernels accomplish.

The identity map $\mathrm{id}_X : X \nrightarrow X$ defined by $(\mathrm{id}_X)(x'|x) = [\![x = x']\!]$, where the Iverson bracket $[\![\cdot]\!]$ evaluates to 1 if $\cdot$ is true and 0 otherwise, is a bare line:

$$\mathrm{id}_X := \; X \text{-} X \tag{11}$$

We choose a particular 1-element set $\{*\}$ that acts as the identity in the sense that $\{*\} \times A \cong A \times \{*\} \cong A$ for any set A. The erase map $\mathrm{del}_X : X \nrightarrow \{*\}$ defined by $(\mathrm{del}_X)(*|x) = 1$ is a Markov kernel that "discards the input". It is drawn as a fuse:

$$\mathrm{del}_X := \; \multimap\!\!* \; X \tag{12}$$

The copy map $\mathrm{copy}_X : X \nrightarrow X \times X$ defined by $(\mathrm{copy}_X)(x', x''|x) = [\![x = x']\!][\![x = x'']\!]$ is a Markov kernel that makes two identical copies of the input. It is drawn as a fork:

$$\mathrm{copy}_X := \; X \multimap \begin{array}{l} X \\ X \end{array} \tag{13}$$

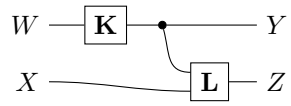
The swap map $\mathrm{swap}_{X,Y} : X \times Y \nrightarrow Y \times X$ defined by $(\mathrm{swap}_{X,Y})(y', x'|x, y) = [\![x = x']\!][\![y = y']\!]$ swaps two inputs, and is represented by crossing wires:

$$\mathrm{swap}_X := \; \times \tag{14}$$

Because we anticipate that the graphical notation will be unfamiliar to many, we will also include translations to more familiar notation.

2.5 Combs: Markov kernels with holes in them

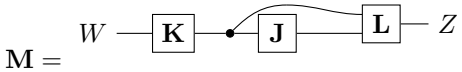
Consider the Markov kernels $\mathbf{K} : W \nrightarrow Y$, $\mathbf{L} : Y \times X \nrightarrow Z$ and $\mathbf{M} : W \times X \nrightarrow Y \times Z$ defined as

$$\mathbf{M} = \quad (15)$$


Following the rules above, we can translate this to ordinary notation by first breaking it down into products and tensor products, and then evaluating these products

$$\begin{aligned}
\mathbf{M}(y,z|w,x) &= (\mathbf{K}\otimes \mathrm{id}_X)[(\mathrm{copy}_Y\otimes \mathrm{id}_X)(\mathrm{id}_Y\otimes \mathbf{L})](y,z|w,x) && (16)\\
&= \sum_{x'\in X, y'\in Y} \mathbf{K}(y'|w)[\![x=x']\!][(\mathrm{copy}_Y\otimes \mathrm{id}_X)(\mathrm{id}_Y\otimes \mathbf{L})](y,z|y',x') && (17)\\
&= \sum_{y'\in Y} \mathbf{K}(y'|w)[(\mathrm{copy}_Y\otimes \mathrm{id}_X)(\mathrm{id}_Y\otimes \mathbf{L})](y,z|y',x) && (18)\\
&= \sum_{y'} \mathbf{K}(y'|w) \sum_{x'\in X, y'', y'''\in Y} [\![y'=y'']\!][\![y'=y''']\!][\![x=x']\!][\![y''=y]\!]\mathbf{L}(z|y''',x')(y,z|y',x') && (19)\\
&= \mathbf{K}(y|w)\mathbf{L}(z|y,x) && (20)
\end{aligned}$$

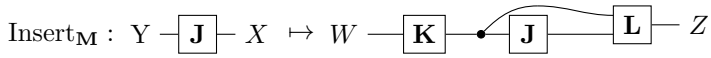
Now, if we are given additionally $\mathbf{J} : Y \nrightarrow X$, we can define a new Markov kernel $\mathbf{N} : W \nrightarrow Z$:

$$\mathbf{M} = \quad (21)$$


This can be translated as

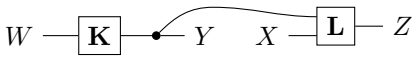
$$\begin{aligned}
\mathbf{N}(z|w) &= [\mathbf{K}\mathrm{copy}_Y(\mathbf{J}\otimes \mathrm{id}_Y)\mathbf{L}](z|w) && (22)\\
&= \sum_{y'\in Y} \mathbf{K}(y'|w) \sum_{y'', y'''\in Y} [\![y'=y'']\!][\![y'=y''']\!][(\mathbf{J}\otimes \mathrm{id}_Y)\mathbf{L}](z|y'',y''') && (23)\\
&= \sum_{y'\in Y} \mathbf{K}(y'|w)[(\mathbf{J}\otimes \mathrm{id}_Y)\mathbf{L}](z|y',y') && (24)\\
&= \sum_{y'\in Y} \mathbf{K}(y'|w) \sum_{x'\in X, y''\in Y} \mathbf{J}(x'|y')[\![y'=y'']\!]\mathbf{L}(z|y'',x') && (25)\\
&= \sum_{y'\in Y, x'\in X} \mathbf{K}(y'|w)\mathbf{J}(x'|y')\mathbf{L}(z|y',x') && (26)\\
& && (27)
\end{aligned}$$

We can define an *insertion* operation $\mathrm{Insert}_{\mathbf{M}} : \Delta(X)^Y \to \Delta(Z)^W$ that takes a Markov kernel $Y \nrightarrow X$ and returns a Markov kernel $W \nrightarrow Z$:

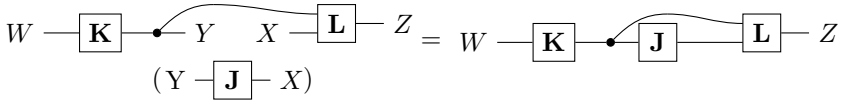


(28)

We can then represent the Insert$_{\mathbf{M}}$ operation (informally) as follows:



(29)



(30)

Thus Insert$_{\mathbf{M}}$ is a "Markov kernel with a hole in it". Such operations are the basis of models with probability holes.

2.6 Probability models with holes

Define $\mathsf{I} : \Omega \to \Omega$ as the identity function. A *standard probability model* on $(\Omega, \mathcal{F})$ is a probability distribution $\mathbb{P}^{\mathsf{I}}$ on $(\Omega, \mathcal{F})$. A *holey probability model* is a function that maps appropriately typed Markov kernels to standard probability models.

A *model* is a collection of conditional probabilities. For example, we could have $\{\mathbb{P}^{\mathsf{X}|*}, \mathbb{P}^{\mathsf{Y}|\mathsf{X}}\}$. We use the convention that conditional probabilities with the same base letter all belong to the same model. Whether a particular model is "correct" is a difficult question to answer, and indeed a difficult question to pose. However, we can specify some necessary conditions that must hold. Informally, these are:

1. All conditional probabilities assign probability 0 to contradictions and probability 1 to tautologies
2. There exists a $\mathsf{Y} : \Omega \to Y$ and a master probability $\mathbb{P}^{\mathsf{I}|\mathsf{Y}}$ compatible with all the conditional probabilities in the model

The first condition is not standard, while the second condition is.

Definition 2.1 (Probability 0 to contradictions)**.** A conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ assigns probability 0 to contradictions if

$$(\mathsf{X}^{-1}(x) \neq \emptyset) \wedge (\mathsf{X}^{-1}(x) \cap \mathsf{Y}^{-1}(A) = \emptyset) \implies \mathbb{P}^{\mathsf{Y}|\mathsf{X}}(A|x) = 0 \tag{31}$$

This condition says: if, according to my chosen variables, X may potentially yield x but Y cannot possibly yield A when X yields x, then the conditional probability must assign a probability of 0 to Y yielding A given X yields x. This is non-standard, because the condition is not that the *probability* that $\mathbb{P}^{\mathsf{X}}(x)$ is nonzero (a quantity which may not be defined), but simply that x is in the range of the variable X. This is required if $\mathbb{P}^{\mathsf{X}}$ is not defined. Hájek (2003) argues that it is desirable even if $\mathbb{P}^{\mathsf{X}}$ is defined, though in that case it would normally have no impact on questions of interest. See also Rényi (1956).

It is a consequence of Definition 2.1 that $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ assigns probability 1 to tautologies. If $\mathsf{X} \bowtie x \subset \mathsf{Y} \bowtie A$ then $\mathsf{X}^{-1}(x) \cap \mathsf{Y}^{-1}(A^C) = \emptyset$, and hence $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}(A^C|x) = 0$ so $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}(A|x) = 1$.

Definition 2.2 (Compatible master condtional probability)**.** We say a conditional probability $\mathbb{P}^{\mathsf{X}_i|\mathsf{X}_j}$ is compatible with $\mathbb{P}^{\mathsf{I}|\mathsf{Y}}$ if

$$\mathbb{P}^{\mathsf{I}|\mathsf{Y}}((\mathsf{X}_i, \mathsf{X}_j) \bowtie A \times B|c) = \int_A \mathbb{P}^{\mathsf{X}_i|\mathsf{X}_j}(B|\mathsf{X}_j(\omega))\mathbb{P}^{\mathsf{I}|\mathsf{Y}}(d\omega|c) \quad \forall c \in Y : \mathbb{P}^{\mathsf{I}|\mathsf{Y}}(\mathsf{X}_j \bowtie A|c) > 0 \tag{32}$$

There is one final difference between our approach and the standard one: we say a model is a collection of conditional probabilities compatible with some $\mathbb{P}^{\mathsf{I}|\mathsf{Y}}$, while the standard approach is to say that $\mathbb{P}^{\mathsf{I}|\mathsf{Y}}$ is the model and goes on to define how the conditional probabilities can be derived from it. The difference here is that **our approach implies that conditional probabilities in a model are unique**. If there are multiple Markov kernels that satisfy Equation 32, we assume that one of these is chosen somehow.

2.7 Composition and probability with variables

We then need a notion of Markov kernels that "maps between variables". An *indexed Markov kernel* is such a thing.

Definition 2.3 (Indexed Markov kernel)**.** Given variables $\mathsf{X} : \Omega \to A$ and $\mathsf{Y} : \Omega \to B$, an indexed Markov kernel $\mathbf{K} : \mathsf{X} \twoheadrightarrow \mathsf{Y}$ is a triple $(\mathbf{K}', \mathsf{X}, \mathsf{Y})$ where $\mathbf{K}' : A \twoheadrightarrow B$ is the *underlying kernel*, X is the *input index* and Y is the *output index*.

For example, if $\mathbf{K} : (\mathsf{A}_1, \mathsf{A}_2) \to \Delta(\mathsf{B}_1, \mathsf{B}_2)$, for example, we can draw:

$$\mathbf{K} := \begin{array}{c}\mathsf{A}_1 \\ \mathsf{A}_2\end{array} \boxed{\mathbf{K}} \begin{array}{c}\mathsf{B}_1 \\ \mathsf{B}_2\end{array} \tag{33}$$

or

$$\mathbf{K} = (\mathsf{A}_1, \mathsf{A}_2) \boxed{\mathbf{K}[\mathbb{L}]} (\mathsf{B}_1, \mathsf{B}_2) \tag{34}$$

We define the product of indexed Markov kenrnels $\mathbf{K} : \mathsf{X} \twoheadrightarrow \mathsf{Y}$ and $\mathbf{L} : \mathsf{Y} \twoheadrightarrow \mathsf{Z}$ as the triple $\mathbf{KL} := (\mathbf{K}'\mathbf{L}', \mathsf{X}, \mathsf{Z})$.

Similarly, the tensor product of $\mathbf{K} : \mathsf{X} \nrightarrow \mathsf{Y}$ and $\mathbf{L} : \mathsf{W} \nrightarrow \mathsf{Z}$ is the triple $\mathbf{K} \otimes \mathbf{L} := (\mathbf{K}' \otimes \mathbf{L}', (\mathsf{X}, \mathsf{W}), (\mathsf{Y}, \mathsf{Z}))$.

We define Id_{X} to be the model $(\mathrm{Id}_X, \mathsf{X}, \mathsf{X})$, and similarly the indexed versions $\mathrm{del}_{\mathsf{X}}$, $\mathrm{copy}_{\mathsf{X}}$ and $\mathrm{swap}_{\mathsf{X},\mathsf{Y}}$ are obtained by taking the unindexed versions of these maps and attaching the appropriate random variables as indices. Diagrams are the diagrams associated with the underlying kernel, with input and output wires annotated with input and output indices.

The category of indexed Markov kernels as morphisms and variables as objects is a Markov category (Appendix 10), and so a valid derivation based on the string diagram language for Markov categories corresponds to a valid theorem in this category. However, most of the diagrams we can form are not viable candidates for models of our variables. For example, if X takes values in $\{0, 1\}$ we can propose an indexed Markov kernel $\mathbf{K} : \mathsf{X} \nrightarrow \mathsf{X}$ with $\mathbf{K}'^{b}_{a} = 0.5$ for all a, b. However, this is not a useful model of the variable X – it expresses something like "if we know the value of X, then we belive that X could take any value with equal probability".

We define a *model* as "an indexed Markov kernel that assigns probability 0 to things known to be contradictions". A contradiction is a simultaneous assignment of values to the variables X and Y such that there is no value of ω under which they jointly take these values. Unless the value assignment to the domain variable is itself contradictory, we hold that any valid model must assign probability zero to such occurrences.

Definition 2.4 (Probabilistic model)**.** An indexed Markov kernel $(\mathbf{K}', \mathsf{X}, \mathsf{Y})$ is a *probabilistic model* ("model" for short) if it is *consistent*, which means it assigns probability 0 to contradictions:

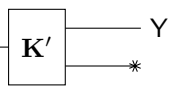
$$f_{\mathsf{X}}^{-1}(a) \cap f_{\mathsf{Y}}^{-1}(b) = \emptyset \implies \left(\mathbf{K}'^{b}_{a} = 0\right) \vee \left(f_{\mathsf{X}}^{-1}(a) = \emptyset\right) \tag{35}$$

A *probability model* is a model where the underlying kernel $\mathbf{K}'$ has the unit I as the domain. We use the font $\mathbb{K}$ to distinguish models from arbitrary indexed Markov kernels.

Consistency implies that for any $\mathbb{K} : \mathsf{X} \nrightarrow \mathsf{Y}$, if $f_{\mathsf{Y}} = g \circ f_{\mathsf{X}}$ then $\mathbb{K}^{g(x)}_{x} = 1$. A particularly simple case of this is a model $\mathbb{L} : \mathsf{X} \nrightarrow \mathsf{X}$, which must be such that $\mathbb{L}^{x}_{x} = 1$. Hájek (2003) has pointed out that standard definitions of conditional probability allow the conditional probability to be arbitrary on a set of measure zero, even though "the probability $\mathsf{X} = x$, given $\mathsf{X} = x$" should obviously be 1.

We take the idea of marginal distributions as fundamental.

Definition 2.5 (Marginal distribution)**.** Given a model $\mathbb{K} : \mathsf{X} \nrightarrow (\mathsf{Y}, \mathsf{Z})$, the marginal distribution of Y, written $\mathbb{K}^{\mathsf{Y}|\mathsf{X}}$, is obtained by marginalising over Z:

$\mathbb{K}^{\mathsf{Y}|\mathsf{X}} :=$ 

(36)

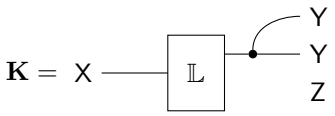
$$\iff \tag{37}$$

$$(\mathbb{K}^{\mathsf{Y}|\mathsf{X}})^{y}_{x} = \sum_{z \in Z} \mathbf{K}'^{yz}_{x} \tag{38}$$

Definition 2.6 (Disintegration)**.** Given a model $\mathbb{K} : \mathsf{X} \nrightarrow (\mathsf{Y}, \mathsf{Z})$, a disintegration $\mathbb{L} : (\mathsf{X}, \mathsf{Y}) \nrightarrow \mathsf{Z}\ \mathsf{Y}$, written $\mathbb{K}^{\mathsf{Y}|\mathsf{X}}$, is obtained by marginalising over Z

We can always get a valid model by adding a copy map to a valid model, and conversely all valid models with repeated codomain variables must contain copy maps.

Lemma 2.7 (Output copies of the same variable are identical)**.** *For any* $\mathbf{K} : \mathsf{X} \nrightarrow (\mathsf{Y}, \mathsf{Y}, \mathsf{Z})$, $\mathbf{K}$ *is a model iff there exists some* $\mathbb{L} : \mathsf{X} \nrightarrow (\mathsf{Y}, \mathsf{Z})$ *such that*



(39)

$$\iff \tag{40}$$

$$\mathbf{K}'^{y,y',z}_{x} = [\![y = y']\!]\mathbf{L}'^{y,z}_{x} \tag{41}$$

$$\tag{42}$$

Proof. $\implies$ For any ω, x, y, y', z:

$$(\mathsf{X}, \mathsf{Y}, \mathsf{Y}, \mathsf{Z})^{x,y,y',z}_{\omega} = [\![f_{\mathsf{Y}}(\omega) = y]\!][\![f_{\mathsf{Y}}(\omega) = y']\!](\mathsf{X}, \mathsf{Z})^{x,z}_{\omega} \tag{43}$$

$$= [\![y = y']\!][\![f_{\mathsf{Y}}(\omega) = y]\!](\mathsf{X}, \mathsf{Z})^{x,z}_{\omega} \tag{44}$$

Therefore, by consistency, for any x, y, y', z, $y \neq y' \implies \mathbf{K}'^{yy'z}_{x} = 0$. Define $\mathbf{L}$ by $\mathbf{L}'^{y,z}_{x} := \mathbf{K}'^{yyz}_{x}$. The fact that $\mathbb{L}$ is a model follows from the assumption that $\mathbb{K}$ is. Then

$$\mathbf{K}'^{y,y',z}_{x} = [\![y = y']\!]\mathbf{L}'^{y,z}_{x} \tag{45}$$

$\Leftarrow$ If $\mathbb{L}$ is a model, then for any x, x', y, z,

$$[\![y = y']\!]\mathbb{L}'^{y,z}_{x} > 0 \implies y = y' \wedge \mathbb{L}'^{y,z}_{x} > 0 \tag{46}$$

$$\implies \left(f_{\mathsf{X}}^{-1}(x) = \emptyset\right) \vee \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) \cap f_{\mathsf{Y}}^{-1}(y) \cap f_{\mathsf{Z}}^{-1}(z) \neq \emptyset\right) \tag{47}$$

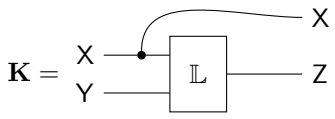
$$\tag{48}$$

□

We can always get a valid model by copying the input to the output of a valid model, and conversely all valid models where there is a variable shared between the input and the output must copy that input to the output.

Lemma 2.8 (Copies shared between input and output are identical)**.** *For any*

$\mathbf{K} : (\mathsf{X}, \mathsf{Y}) \nrightarrow (\mathsf{X}, \mathsf{Z})$, $\mathbf{K}$ *is a model iff there exists some* $\mathbb{L} : (\mathsf{X}, \mathsf{Y}) \nrightarrow \mathsf{Z}$ *such that*

$\mathbf{K} =$ 

(49)

$$\iff \tag{50}$$

$$\mathbf{K}^{\prime x', z}_{x, y} = [\![x = x']\!] \mathbf{L}^{z}_{\prime x, y} \tag{51}$$

Proof. $\implies$ For any ω, x, y, y', z:

$$(\mathsf{X}, \mathsf{Y}, \mathsf{Y}, \mathsf{Z})^{x, y, y', z}_{\omega} = [\![f_{\mathsf{Y}}(\omega) = y]\!][\![f_{\mathsf{Y}}(\omega) = y']\!](\mathsf{X}, \mathsf{Z})^{x, z}_{\omega} \tag{52}$$

$$= [\![y = y']\!][\![f_{\mathsf{Y}}(\omega) = y]\!](\mathsf{X}, \mathsf{Z})^{x, z}_{\omega} \tag{53}$$

Therefore, by consistency, for any x, y, y', z, $x \neq x' \implies \mathbb{K}^{\prime x' z}_{x, y} = 0$. Define $\mathbf{L}$ by $\mathbf{L}^{\prime x', z}_{x, y} := \mathbb{K}^{\prime x, y}_{x, y}$. The fact that $\mathbf{L}$ is a model follows from the assumption that $\mathbb{K}$ is a model. Then

$$\mathbf{K}^{\prime x', z}_{x, y} = [\![x = x']\!] \mathbf{L}^{\prime z}_{x, y} \tag{54}$$

$\impliedby$ If $\mathbb{L}$ is a model, then for any x, x', y, z,

$$[\![x = x']\!] \mathbb{L}^{\prime z}_{x, y} > 0 \implies x = x' \wedge \mathbb{L}^{\prime z}_{x, y} > 0 \tag{55}$$

$$\implies \left(f^{-1}_{\mathsf{X}}(x) \cap f^{-1}_{\mathsf{Y}}(y) = \emptyset\right) \vee \left(f^{-1}_{\mathsf{X}}(x) \cap f^{-1}_{\mathsf{X}}(x) \cap f^{-1}_{\mathsf{Y}}(y) \cap f^{-1}_{\mathsf{Z}}(z) \neq \emptyset\right) \tag{56}$$

$$\tag{57}$$

□

Consistency along with the notion of marginal distributions implies that, given some X and some $\mathbb{K} : \mathsf{Y} \nrightarrow \mathrm{Id}_{\Omega}$, the pushforward $\mathbb{K}\mathbb{X}$ is the unique model $\mathsf{Y} \nrightarrow \mathsf{X}$ that can be paired (Definition 2.10) with $\mathbb{K}$. This is shown in Lemma 2.11.

Lemma 2.9 (Uniqueness of models with the sample space as a domain)**.** *For any* $\mathsf{X} : \Omega \to A$*, there is a unique model* $\mathbb{X} : Id_{\Omega} \nrightarrow \mathsf{X}$ *given by* $\mathbb{X} := (\mathsf{X}, Id_{\Omega}, \mathsf{X})$*.*

Proof. X is a Markov kernel mapping from $\Omega \to A$, so it is a valid underlying kernel for $\mathbb{X}$, and $\mathbb{X}$ has input and output indices matching its signature. We need to show it satisfies consistency.

For any $\omega \in \Omega$, $a \in A$

$$\max_{\omega \in \Omega} (\mathrm{Id}_{\Omega}, \mathsf{X})^{\omega', a}_{\omega} = \max_{\omega \in \Omega} [\![\omega = \omega']\!][\![\omega = f_{\mathsf{X}}(a)]\!] \tag{58}$$

$$= [\![\omega = f_{\mathsf{X}}(a)]\!] \tag{59}$$

$$= \mathbf{X}^{a}_{\omega} \tag{60}$$

Thus $\mathbb{X}$ satisfies consistency.

Suppose there were some $\mathbb{K} : \mathrm{Id}_\Omega \nrightarrow \mathsf{X}$ not equal to X. Then there must be some $\omega \in \Omega$, $b \in A$ such that $\mathbb{K}^b_\omega \neq 0$ and $f_{\mathsf{X}}(\omega) \neq b$. Then

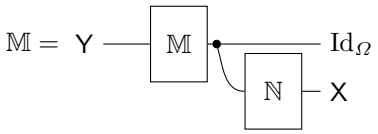
$$\begin{aligned}
\max_{\omega \in \Omega} (\mathrm{Id}_\Omega, \mathsf{X})^{\omega', a}_{\omega} &= \max_{\omega \in \Omega} [\![\omega = \omega']\!] [\![\omega = f_{\mathsf{X}}(b)]\!] && (61)\\
&= [\![\omega = f_{\mathsf{X}}(b)]\!] && (62)\\
&= 0 && (63)\\
&< \mathbb{K}^b_\omega && (64)
\end{aligned}$$

Thus $\mathbb{K}$ doesn't satisfy consistency. □

Definition 2.10 (Pairing)**.** Two models $\mathbb{K} : \mathsf{X} \nrightarrow \mathsf{Y}$ and $\mathbb{L} : \mathsf{X} \nrightarrow \mathsf{Z}$ can be *paired* if there is some $\mathbb{M} : \mathsf{X} \nrightarrow (\mathsf{Y}, \mathsf{Z})$ such that $\mathbb{K} = \mathbb{M}^{\mathsf{Y}|\mathsf{X}}$ and $\mathbb{L} = \mathbb{M}^{\mathsf{Z}|\mathsf{X}}$.

Lemma 2.11 (Pushforward model)**.** *Given any model $\mathbb{K} : \mathsf{Y} \nrightarrow Id_\Omega$ and any X, there is a unique $\mathbb{L} : \mathsf{Y} \nrightarrow \mathsf{X}$ that can be paired with $\mathbb{K}$, and it is given by $(\mathbf{L}^a_b = \sum_{\omega \in f_{\mathsf{X}}^{-1}(a)} \mathbf{K}^\omega_b$.*

Proof. Suppose that there is some $\mathbb{L}$ that can be paired with $\mathbb{K}$ via some $\mathbb{M} : \mathsf{Y} \nrightarrow (\mathrm{Id}_\Omega, \mathsf{X})$. Then, by the existence of disintegrations, there must be some $\mathbb{N} : \mathrm{Id}_\Omega \nrightarrow \mathsf{X}$ such that

$\mathbb{M} =$ 

(65)

By Corollary **??**, there is only one model $\mathbb{N} : \mathrm{Id}_\Omega \nrightarrow \mathsf{X}$ is unique and equal to $\mathbb{X} := (\mathsf{X}, \mathrm{Id}_\Omega, \mathsf{X})$.

It remains to be shown that $\mathbb{M}$ is also a model. We already know that $\mathbb{K}$ is consistent with respect to $(\mathsf{Y}, \mathrm{Id}_\Omega)$ and $\mathbb{L}$ is consistent with respect to $(\mathrm{Id}_\Omega, \mathsf{X})$. $\mathbb{M}$ must be consistent with respect to $(\mathsf{Y}, \mathrm{Id}_\Omega, \mathsf{X})$. Consider any $x \in X$, $\omega \in \Omega$, $y \in Y$ such that $f^{-1}_{\mathsf{X}}(x) \cap \{\omega\} \neq \emptyset$ and $f^{-1}_{\mathsf{Y}}(y) \cap \{\omega\} \neq \emptyset$. Trouble might arise if $f^{-1}_{\mathsf{X}}(x) \cap \{\omega\} \cap f^{-1}_{\mathsf{Y}}(y) = \emptyset$, but this is obviously impossible as $\omega \in f^{-1}_{\mathsf{X}}(x)$ and $\omega \in f^{-1}_{\mathsf{Y}}(y)$.

Finally, for any $a \in A$, $b \in B$

$$\begin{aligned}
(\mathbb{K}\mathbb{X})^a_b &= \sum_{\omega \in \Omega} \mathbb{P}^\omega_b \mathsf{X}^a_\omega && (66)\\
&= \sum_{\omega \in \Omega} \mathbb{P}^\omega_b [\![a = f_{\mathsf{X}}(\omega)]\!] && (67)\\
&= \sum_{\omega \in f^{-1}(a)} \mathbb{P}^\omega_b && (68)
\end{aligned}$$

□

Corollary 2.12 (Pushforward probability model)**.** *Given any probability model* $\mathbb{P} : \mathsf{I} \to Id_\Omega$*, there is a unique model* $\mathbb{P}^{\mathsf{X}} : \mathsf{I} \to \mathsf{X}$ *such that* $\mathbb{P}^{\mathsf{X}} = \mathbb{P}\mathbb{Q}$ *for some* $\mathbb{Q} : Id_\Omega \to \mathsf{X}$*, and it is given by* $(\mathbb{P}^{\mathsf{X}})^a_b = \sum_{\omega \in f^{-1}(a)} \mathbb{P}^\omega_b$*.*

Proof. Apply Lemma 2.11 to a model $\mathbb{P} : \mathsf{I} \to \mathrm{Id}_\Omega$. □

The following lemmas can help us check whether an indexed Markov kernel is a valid model.

We take the following term from Constantinou and Dawid (2017). Our definition is equivalent to unconditional variation independence in that paper.

Definition 2.13 (Variation independence)**.** Two variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$ are variation independent, written $\mathsf{X} \perp_v \mathsf{Y}$, if for all $y \in f_{\mathsf{Y}}(\Omega) R (f_{\mathsf{Y}})$

$$f_{\mathsf{Y}}(\Omega) \times f_{\mathsf{X}}(\Omega) = \{(f_{\mathsf{Y}}(\omega), f_{\mathsf{X}}(\omega)) | \omega \in \Omega\} \tag{69}$$

If a collection of variables is variation independent and surjective, then an arbitrary indexed Markov kernel labelled with these variables is a model.

Lemma 2.14 (Consistency via variation conditional independence)**.** *Given an indexed Markov kernel* $\mathbf{K} : \mathsf{X} \to \mathsf{Y}$ *with* $\mathsf{X} : \Omega \to X$ *and* $\mathsf{Y} : \Omega \to Y$*, if* f_{Y} *is surjective and* $\mathsf{Y} \perp_v \mathsf{X}$ *then* $\mathbf{K}$ *is a model.*

Proof. By variation independence and surjectivity of f_{Y}, for any $x \in X$, $y \in Y$, $f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) = \emptyset \implies f_{\mathsf{X}}^{-1}(x) = \emptyset$. Thus the criterion of consistency places no restrictions on $\mathbf{K}$. □

I think Lemmas 2.7 and 2.8 might be sufficient to offer diagrammatic checks of consistency if all variables that are not identical are variation independent. This is probably an interesting result, but I'm not sure if it's a higher priority than filling out the rest of the content.

Alternatively, if we have a strictly positive indexed Markov kernel that is known to be a model, we can conclude that arbitrary indexed Markov kernels with appropriate labels are also models.

Lemma 2.15 (Consistency via positive models)**.** *Given a model* $\mathbb{K} : \mathsf{X} \to (\mathsf{Y}, \mathsf{Z})$*, if an indexed Markov kernel* $\mathbf{L} : (\mathsf{X}, \mathsf{Y}) \to \mathsf{Z}$ *has the property* $\mathbf{K}'^{yz}_x = 0 \implies \mathbf{L}'^{z}_{xy} = 0$ *then* $\mathbf{L}$ *is also a model.*

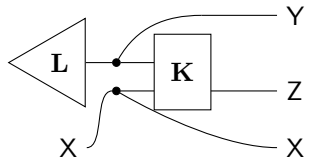
Proof. Because $\mathbb{K}$ is a model,

$$\begin{aligned}
\mathbf{L}'^{z}_{xy} > 0 &\implies \mathbf{K}'^{yz}_{x} > 0 && (70)\\
&\implies \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) \cap f_{\mathsf{Z}}^{-1}(z) \neq \emptyset\right) \vee \left(f_{\mathsf{X}}^{-1}(x) = \emptyset\right) && (71)\\
&\implies \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) \cap f_{\mathsf{Z}}^{-1}(z) \neq \emptyset\right) \vee \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) = \emptyset\right) && (72)
\end{aligned}$$

□

2.8 Truncated factorisation with variables

At this point, we can represent Equation 6 using models. Suppose $P^{\mathsf{Y}|\mathsf{XZ}}$ is an model $\mathbb{K} : (\mathsf{X}, \mathsf{Z}) \nrightarrow \mathsf{Y}$ and $\mathbb{P}^{\mathsf{Z}}$ an model $\mathbb{L} : \{*\} \nrightarrow \mathsf{Z}$. Then we can define an indexed Markov kernel $\mathbf{M} : \mathsf{X} \nrightarrow \mathsf{X}, \mathsf{Z}$ representing $x \mapsto \mathbb{P}_x^{\mathsf{YZ}}(y, z)$ by

$\mathbf{M} :=$ 

(73)

Equation 73 is almost identical to Equation **??**, except it now specifies which variables each measure applies to, not just which sets they take values in. Like the original Equation 6, there is no guarantee that $\mathbf{M}$ is actually a model. If $f_{\mathsf{X}} = g \circ f_{\mathsf{Z}}$ for some $g : Z \to X$ and X has more than 1 element, then the rule of consistency will rule out the existence of any such model.

If we want to use $\mathbf{M}$, we want it at minimum to satisfy the consistency condition. One approach we could use is to check the result using Lemmas 2.7 to 2.15, although note that 2.14 and 2.15 are sufficient conditions, not necessary ones.

2.9 Sample space models and submodels

Instead of trying to assemble probability models as in Equation 73, we might try to build probability models in a manner closer to the standard setup – that is, we start with a sample space model (or a collection of sample space models) and work with marginal and conditional probabilities derived from these, without using any non-standard model assemblies.

A sample space model is any model $\mathbf{K} : \mathsf{X} \nrightarrow \mathrm{Id}_\Omega$. We expect that the collection of models under consideration will usually be defined on some small collection of random variables, but every such model is the pushforward of some sample space model. Using sample space models allows us to stay close to the usual convention of probability modelling that starts with a sample space probability model.

Lemma 2.16 (Existence of sample space model)**.** *Given any model $\mathbb{K} : \mathsf{X} \nrightarrow \mathsf{Y}$, there is a sample space model $\mathbb{L} : \mathsf{X} \nrightarrow Id_\Omega$ such that, defining $\mathbb{Y} := (\mathsf{Y}, Id_\Omega, \mathsf{Y})$, $\mathbb{L}\mathbb{Y} = \mathbb{K}$.*

Proof. If $\mathsf{X} : \Omega \nrightarrow A$ and $\mathsf{Y} : \Omega \nrightarrow B$, take any $a \in A$ and $b \in B$. Then set

$$\mathbf{L}'^{\omega}_a = \begin{cases} 0 & \text{if } f_{\mathsf{Y}}^{-1}(b) \cap f_{\mathsf{X}}^{-1}(a) = \emptyset \\ \mathbf{K}'^{b}_a[\![\omega = \omega_b]\!] & \text{for some } \omega_b \in f_{\mathsf{Y}}^{-1}(b) \text{ if } f_{\mathsf{X}}^{-1}(a) = \emptyset \\ \mathbf{K}'^{b}_a[\![\omega = \omega_{ab}]\!] & \text{for some } \omega_{ab} \in f_{\mathsf{Y}}^{-1}(b) \cap f_{\mathsf{X}}^{-1}(a) \text{ otherwise} \end{cases} \tag{74}$$

Note that for all $a \in A$, $\sum_{\omega\in\Omega} \mathbf{L}'^{\omega}_a = \sum_{b\in B} \mathbf{K}'^{b}_a = 1$.
By construction, $(\mathbf{L}', \mathrm{Id}_\Omega, \mathsf{X})$ is free of contradiction. In addition

$$(\mathbf{L}'\mathsf{Y})^b_a = \sum_{\omega\in\Omega} \mathbf{L}'^{\omega}_a \mathsf{Y}^b_\omega \tag{75}$$

$$= \sum_{\omega\in f_{\mathsf{Y}}^{-1}(b)} \mathbf{L}'^{\omega}_a \tag{76}$$

$$= \begin{cases} 0 & f_{\mathsf{Y}}^{-1}(b) \cap f_{\mathsf{X}}^{-1}(a) = \emptyset \\ \mathbf{K}'^{b}_a & \text{otherwise} \end{cases} \tag{77}$$

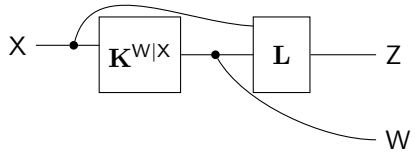
$$\implies (\mathbf{L}'\mathsf{Y}) = \mathbf{K}' \tag{78}$$

□

Definition 2.17 (Pushforward model)**.** For any variables $\mathsf{X} : \Omega \twoheadrightarrow A$, $\mathsf{Y} : \Omega \twoheadrightarrow B$ and any sample space model $\mathbb{K} : \mathsf{X} \twoheadrightarrow \mathrm{Id}_\Omega$, the pushforward $\mathbb{K}^{\mathsf{Y}|\mathsf{X}} := \mathbb{K}\mathbb{X}$ where $\mathbb{X} := (\mathsf{X}, \mathrm{Id}_\Omega, \mathsf{X})$.

The fact that the pushforward is a model is proved in Lemma 2.11. We employ the slightly more familiar notation $\mathbb{K}^{\mathsf{Y}|\mathsf{X}}(y|x) \equiv (\mathbf{K}'^{\mathsf{Y}|\mathsf{X}})^y_x$.

Definition 2.18 (Submodel)**.** Given $\mathbb{K} : \mathsf{X} \twoheadrightarrow \mathrm{Id}_\Omega$ and $\mathbb{L} : \mathsf{W}, \mathsf{X} \twoheadrightarrow \mathsf{Z}$, $\mathbb{L}$ is a submodel of $\mathbb{K}$ if

$\mathbb{K}^{\mathsf{Z},\mathsf{W}|\mathsf{Y}} =$ 

(79)

$$(\mathbb{K}^{\mathsf{Z},\mathsf{W}|\mathsf{Y}})^{w,z}_x = (\mathbb{K}^{\mathsf{W}|\mathsf{Y}})^w_x \mathbb{L}^z_{w,x} \tag{80}$$

We write $\mathbb{L} \in \mathbb{K}^{\{\mathsf{Z}|\mathsf{W},\mathsf{X}\}}$.

Lemma 2.19 (Submodel existence)**.** *For any model $\mathbb{K} : \mathsf{X} \twoheadrightarrow \mathrm{Id}_\Omega$ (where Ω is a finite set), W and Y, there exists a submodel $\mathbb{L} : (\mathsf{W}, \mathsf{X}) \twoheadrightarrow \mathsf{Y}$.*

Proof. Consider any indexed Markov kernel $\mathbf{L} : (\mathsf{W}, \mathsf{X}) \twoheadrightarrow \mathsf{Y}$ with the property

$$\mathbf{L}'^{y}_{wx} = \frac{\mathbb{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w, y|x)}{\mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x)} \qquad \forall x, w : \text{ the denominator is positive} \tag{81}$$

In general there are many indexed Markov kernels that satisfy this. We need to check that $\mathbf{L}'$ can be chosen so that it avoids contradictions. For all x, y such that $\mathbf{K}^{\mathsf{Y}|\mathsf{X}}(y|x)$ is positive, we have $\mathbb{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w, y|x) > 0 \implies \mathbf{L}'^{y}_{wx} > 0$. Furthermore, where $\mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x) = 0$, we either have $f_{\mathsf{W}}^{-1}(w) \cap f_{\mathsf{X}}^{-1}(x) = \emptyset$ or we can choose some $\omega_{wx} \in f_{\mathsf{W}}^{-1}(w) \cap f_{\mathsf{X}}^{-1}(x)$ and let $\mathbf{L}'^{f_{\mathsf{Y}}(\omega_{wx})}_{wx} = 1$. Thus $\mathbf{L}'$ can be chosen such that $\mathbf{L}$ is a model (but this is not automatic).

Then

$$\mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x)\mathbf{L}'^{y}_{xw} = \mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x)\frac{\mathbb{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w,y|x)}{\mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x)} \qquad \text{if } \mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x) > 0 \tag{82}$$

$$= \mathbb{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w,y|x) \qquad \text{if } \mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x) > 0 \tag{83}$$

$$= 0 \qquad \text{otherwise} \tag{84}$$

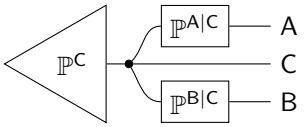
$$= \mathbb{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w,y|x) \qquad \text{otherwise} \tag{85}$$

□

2.10 Conditional independence

We define conditional independence in the following manner:

For a *probability model* $\mathbb{P} : \mathsf{I} \nrightarrow \mathrm{Id}_\Omega$ and variables $(\mathsf{A}, \mathsf{B}, \mathsf{C})$, we say A is independent of B given C, written $\mathsf{A} \perp\!\!\!\perp_{\mathbb{P}} \mathsf{B}|\mathsf{C}$, if

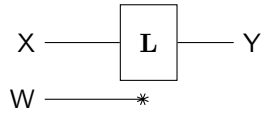
$\mathbf{P}^{\mathsf{ABC}} =$ 

(86)

For an arbitrary model $\mathbf{N} : \mathsf{X} \nrightarrow \mathrm{Id}_\Omega$ where $\mathsf{X} : \Omega \nrightarrow X$, and some $(\mathsf{A}, \mathsf{B}, \mathsf{C})$, we say A is independent of B given C, written $\mathsf{A} \perp\!\!\!\perp_{\mathbf{N}} \mathsf{B}|\mathsf{C}$, if there is some $\mathbb{O} : \mathsf{I} \nrightarrow \mathsf{X}$ such that $O^x > 0$ for all $x \in f_{\mathsf{X}}^{-1}(X)$ and $\mathsf{A} \perp\!\!\!\perp_{\mathbb{ON}} \mathsf{B}|\mathsf{C}$.

This definition is inappliccable in the case where sets may be uncountably infinite, as no such $\mathbf{O}$ can exist in this case. There may well be definitions of conditional independence that generalise better, and we refer to the discussions in Fritz (2020) and Constantinou and Dawid (2017) for some discussion of alternative definitions. One advantage of this definition is that it matches the version given by Cho and Jacobs (2019) which they showed coincides with the standard notion of conditional independence and so we don't have to show this in our particular case.

A particular case of interest is when a kernel $\mathbf{K} : (\mathsf{X}, \mathsf{W}) \to \Delta(\mathsf{Y})$ can, for some $\mathbf{L} : \mathsf{W} \to \Delta(\mathsf{Y})$, be written:

$\mathbf{K} =$ 

(87)

Then $\mathsf{Y} \perp\!\!\!\perp_{\mathbf{K}} \mathsf{W}|\mathsf{X}$.

3 Decision theoretic causal inference

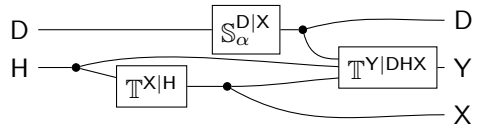
The first question we want to investigate is: supposing that we are happy to use the modelling approach described in the previous section, what kind of model would we want to use to help make good choices when we have to make choices?

Suppose we will be given an observation, modelled by X taking values in X, and in response to this we can select any decision, modelled by D taking values in D. The process by which we choose a decision or mixture of decisions, is called a decision rule or a *strategy*, designated α and modelled by $\mathbf{S}_\alpha : \mathsf{X} \to \Delta(\mathsf{D})$[2]. We assume that the collection of strategies under consideration $\{\mathbf{S}_\alpha\}_\alpha$ is convex. We are interested in some defined collection of things that will be determined at some point after we have taken our decision; these will be modelled by the variable Y and we will call them *consequences*.

For different observations and decisions we will generally expect different consequences. We will assume that we expect the same observations whatever strategy we choose. We will also assume that given the same observations and the same decision, we expect the same consequences regardless of the strategy. These assumptions rule out certain classes of decision problem where, for example, there is controversy over whether the strategy chosen should depend on the time at which it is chosen Weirich (2016); Lewis (1981); Paul F. Christiano (2018).

We will entertain a collection of probabilistic models to represent postulated relationships between X, D and Y for each strategy α; to do this, we will introduce a latent variable H such that each value of H corresponds to a particular probabilistic model of X, D and Y. Concretely, for each strategy α our forecast will be represented by a probability model $\mathbf{P}_\alpha : \mathsf{I} \twoheadrightarrow (\mathsf{H}, \mathsf{X}, \mathsf{D}, \mathsf{Y})$. We assume that – holding the hypothesis fixed – the same observations are expected whatever strategy we choose: $\mathbb{P}_\alpha^{\mathsf{X}|\mathsf{H}} = P_\beta^{\mathsf{X}|\mathsf{H}}$ for all α, β. We assume that under each hypothesis, the decision chosen is always modelled by the the chosen strategy: $\mathbb{P}_\alpha^{\mathsf{D}|\mathsf{HX}} = \mathbf{S}_\alpha \otimes \text{erase}_{\mathsf{H}}$. Finally, we assume that, holding the hypothesis fixed, the same consequences are expected under any strategy given the same observations and the same decision: $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{XHD}} = P_\beta^{\mathsf{Y}|\mathsf{XHD}}$ for all α, β.

Under these assumptions, there exists a "see-do model" $\mathbb{T}^{\mathsf{XY}|\mathsf{HD}}$ such that $\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$ and for all α,

$\mathbb{P}_\alpha =$ 

(88)

The proof is given in Appendix 6. Note that $\mathbb{T}^{\mathsf{X}|\mathsf{H}}$ exists by virtue of the fact $\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$.

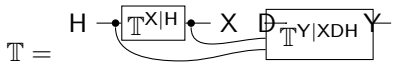
We will call the the see-do model along with the collection of strategies $\{\mathbb{T}^{\mathsf{XY}|\mathsf{HD}}, \{\mathbb{S}_\alpha | \alpha \in \mathcal{A}\}\}$ a *standard decision problem*.

[2]We don't make the strategy a variable simply because we would need an uncountable version of our theory to do it.

3.1 Combs

The conditional independence $\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$ of $\mathbb{T}$ is the property that allows us to write Equation 88, but it also implies that $\mathbb{T}$ is *not* a submodel of $\mathbb{P}_\alpha$ for most strategies α, because for most such strategies X and D are not independent. Instead, $\mathbb{T}$ is a *comb*. This structure was introduced by Chiribella et al. (2008) in the context of quantum circuit architecture, and Jacobs et al. (2019) adapted the concept to causal modelling.

We don't formally define any special operations with combs here, but because they come up multiple times we will explain the notion a little. A comb is a Markov kernel with an "insert" operation; to obtain the probability model associated with a particular strategy, we "insert" the strategy into our see-do model.

$\mathbb{T} =$ 

(89)

$=$ 

(90)

A key feature of a comb is that a strategy can be chosen such that D is independent of any variable on the "upper arm" (X in this example) conditional on H. There is an intuitive appeal to the notion that, with access to a randomiser, we could if we wanted to choose a decision independent of all of our observations. We may wish to introduce additional variables that we do not observe, but we can nonetheless choose D independent of them. Such variables we will call *pre-choice variables*

Definition 3.1 (Pre-choice variable)**.** Given a see-do model $\mathbb{T}$, W is a pre-choice variable iff for every other pre-choice variable V, $(\mathsf{W}, \mathsf{V}) \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$. The hypothesis H is always a pre-choice variable, and we also assume the same is true of the observation X.

Given that H is necessarily a pre-choice variable, we wonder if it may be possible to define a hypothesis H such that all pre-choice variables are functions of it. This would reduce the number of different elements of our theory, as we would no longer distinguish between "hypotheses" and "pre-choice variables". The reason why we have not done so thus far is that hypotheses are motivated by classical statistics while pre-choice variables are motivated by approaches to causal inference, and we haven't yet investigated whether the two can be identified without losing anything important.

Lattimore and Rohde (2019a) and Lattimore and Rohde (2019b) describe a novel approach to causal inference: they consider an observational probability model and a collection of indexed interventional probability models, with the probability model tied to the interventional models by shared parameters. In these papers, they show how such a model can reproduce inferences made using Causal Bayesian Networks. This kind of model is very close to a type of see-do

model, where we identify the hypotheses H with the parameter variables in that work. The only difference is that we consider interventional maps (see-do models represent a map $(\mathsf{D}, \mathsf{H}) \twoheadrightarrow \mathsf{Y}$) rather than interventional probability models, and this is a superficial difference as an indexed collection of probability models is a map.

Dawid (2020) describes a different version of a decision theoretic approach to causal inference:

> A fundamental feature of the DT approach is its consideration of the relationships between the various probability distributions that govern different regimes of interest. As a very simple example, suppose that we have a binary treatment variable T, and a response variable Y. We consider three different regimes [...] the first two regimes may be described as interventional, and the last as observational.

This is somewhat different to a see-do model, as it features a probabilistic model that uses the same random variables T and Y to represent both interventional and observational regimes, while a see-do model uses different random variables. This difference can be thought of as the difference between positing a sequence $(\mathsf{X}_1, \mathsf{X}_2, \mathsf{X}_3)$ distributed according to $\mathbb{P}^{\mathsf{X}}$, or saying that the X_i are distributed according to $\mathbb{P}$ such that they are mutually independent ($i \notin A \subset [3] \implies \mathsf{X}_i \perp\!\!\!\perp_{\mathbb{P}} (\mathsf{X}_j)_{j \in A}$) and identically distributed ($\mathbb{P}^{\mathsf{X}_i} = \mathbb{P}^{\mathsf{X}_j}$ for all i, j). The former can be understood as a shorthand of the latter, but because in this paper we are particularly interested in problems that arise regarding the relation between the map and the territory, we favour the second approach because it is more explicit.

Jacobs et al. (2019) has used a comb decomposition theorem to prove a sufficient identification condition similar to the identification condition given by Tian and Pearl (2002). This theorem depends on the particular inductive hypotheses made by causal Bayesian networks.

3.2 See-do models and classical statistics

See-do models are capable of expressing the expected results of a particular choice of decision strategy, but they cannot by themselves tell us which strategies are more desirable than others. To do this, we need some measure of the desirability of our collection of results $\{\mathbb{P}_\alpha | \alpha \in A\}$. A common way to do this is to employ the principle of expected utility. The classic result of Von Neumann and Morgenstern (1944) shows that all preferences over a collection of probability models that obey their axioms of completeness, transitivity, continuity and independence of irrelevant alternatives must be able to be expressed via the principle of expected utility. This does not imply that anyone knows what the appropriate utility function is.

We introduced the hypothesis H as a latent variable to allow us to postulate multiple different models of obsevations, decisions and consequences. In general, both the hypothesis and the observation X may influence our views about the

consequences Y that are likely to follow from a given decision. It is very common to model sequences of observations as independent and identically distributed given some parameter or latent variable. In such cases, we can identify H with this latent variable (our setup does not preclude introducing a prior over H, nor does it require it). Furthermore, in such cases where we have a collection of X_i such that $\mathsf{X}_i \perp\!\!\!\perp_{\mathbb{T}} \mathsf{X}_j|\mathsf{H}$, it may be reasonable to expect that $\mathsf{Y} \perp\!\!\!\perp_{\mathsf{T}} \mathsf{X}|\mathsf{H}$ also. In fact, this is the standard view in causal modelling – given "the probability distribution over observations" (which is to say, conditional on H), interventional distributions have no additional dependence on *particular* observations. We can find exceptions with questions like "given what actually happened, what would have happened if a different action had been taken?" (Pearl, 2009; Tian and Pearl, 2000; Mueller et al., 2021), but this is not the kind of question we are considering here.

Given these two choices – to use the principle of expected utility to evaluate strategies, and to use a see-do model $\mathbb{T}$ with the conditional independence $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{X}|\mathsf{H}, \mathsf{D}$ – we obtain a statistical decision problem in the form introduced by Wald (1950).

A *statistical model* (or *statistical experiment*) is a collection of probability distributions $\{\mathbb{P}_\theta\}$ indexed by some set Θ. A statistical decision problem gives us an observation variable $\mathsf{X} : \Omega \to\!\!\!\!\bullet\ X$ and a statistical experiment $\{\mathbb{P}^{\mathsf{X}}_\theta\}_\Theta$, a decision set D and a loss $l : \Theta \times D \to \mathbb{R}$. A strategy $\mathbb{S}^{\mathsf{D}|\mathsf{X}}_\alpha$ is evaluated according to the risk functional $R(\theta, \alpha) := \sum_{x\in X}\sum_{d\in D} \mathbb{P}^{\mathsf{X}}_\theta(x) S^{\mathsf{D}|\mathsf{X}}_\alpha(d|x) l(h, d)$. A strategy $\mathbb{S}^{\mathsf{D}|\mathsf{X}}_\alpha$ is considered more desirable than $\mathbb{S}^{\mathsf{D}|\mathsf{X}}_\beta$ if $R(\theta, \alpha) < R(\theta, \beta)$.

Suppose we have a see-do model $\mathbb{T}^{\mathsf{XY}|\mathsf{HD}}$ with $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{X}|(\mathsf{H}, \mathsf{D})$, and suppose that the random variable Y is a "reverse utility" function taking values in $\mathbb{R}$ for which low values are considered desirable. Then, defining a loss $l : H \times D \to \mathbb{R}$ by $l(h, d) = \sum_{y\in\mathbb{R}} y\mathbb{T}^{\mathsf{Y}|\mathsf{HD}}(y|h, d)$, we have

$$\mathbb{E}_{\mathbb{P}_\alpha}[\mathsf{Y}|h] = \sum_{x\in X}\sum_{d\in D}\sum_{y\in Y} \mathbb{T}^{\mathsf{X}|\mathsf{H}}(x|h)\mathbb{S}^{\mathsf{D}|\mathsf{X}}_\alpha(d|x)\mathbb{T}^{\mathsf{Y}|\mathsf{HD}}(y|h, d) \tag{91}$$

$$= \sum_{x\in X}\sum_{d\in D} \mathbb{T}^{\mathsf{X}|\mathsf{H}}(x|h)\mathbb{S}^{\mathsf{D}|\mathsf{X}}_\alpha(d|x) l(h, d) \tag{92}$$

$$= R(h, \alpha) \tag{93}$$

If we are given a see-do model where we interpret $\mathbb{T}^{\mathsf{X}|\mathsf{H}}$ as a statistical experiment and Y as a reversed utility, the expectation of the utility under the strategy forecast given in equation 88 is the risk of that strategy under hypothesis h.

4 Causal Bayesian Networks

When do causal relationships as defined by causal Bayesian networks exist? We will consider a simplified case where a single node may be intervened on, and find the implied see-do model. With this condition, according to Pearl (2009), a

causal Bayesian network is a probability model $\mathbb{P}$, a collection of interventional probability models $\{\mathbb{P}_{\mathsf{X}=a} | a \in X_i\}$ and a directed acyclic graph $\mathcal{G}$ whose nodes are identified with some collection of variables, which we can group into three variables $\{\mathsf{W}, \mathsf{X}, \mathsf{Y}\}$, where W is the sequence of variables associated with the parents of X in $\mathcal{G}$, X is the "intervenable" node of $\mathcal{G}$ and Y are associated with the other nodes. The interventional probability models must all obey the truncated factorisation condition with respect to $\mathcal{G}$:

$$\mathbb{P}^{\mathsf{WXY}}_{\mathsf{X}=a}(w, x, y) = \mathbb{P}^{\mathsf{W}}(w)\mathbb{P}^{\mathsf{Y}|\mathsf{XW}}(y|x, w)[\![x = a]\!] \tag{94}$$

A standard interpretation of the observational and interventional probability distributions is that we have a sequence of observations modeled by $\mathsf{V}_A := (\mathsf{W}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in A}$ mutually independent and identically distributed according to $\mathbb{P}^{\mathsf{WXY}}$, and a sequence of consequences modeled by $\mathsf{V}_B := (\mathsf{W}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i\in B}$ mutually independent and identically distributed according to $\mathbb{P}^{\mathsf{WXY}}_{\mathsf{X}=a}$, and $\mathbb{P}$ and $\mathbb{P}_{\mathsf{X}=a}$ are coupled by Equation 94. What it means for $\mathbb{P}$ and $\mathbb{P}_{\mathsf{X}=a}$ to be coupled is: if $\mathbb{P}$ is the "actual" distribution of observations, then $\mathbb{P}_{\mathsf{X}=a}$ is the "actual" distribution of consequences. This can be explicitly represented by introducing a variable H representing the "actual" distribution of observations, and we introduce a model $\mathbb{U}^{\cdot|\mathsf{H}}$ such that

$$\mathbb{P}^{\mathsf{V}_i} := \mathbb{U}^{\mathsf{V}_i|\mathsf{H}}(v|h) \text{ for some } h \in H \text{ and any } i \in A, v \in W \times X \times Y \tag{95}$$

$$\mathbb{P}^{\mathsf{V}_j}_{\mathsf{X}_j=a} := \mathbb{U}^{\mathsf{V}_j|\mathsf{HX}_j}(v|h, a) \text{ for some } h \in H \text{ and any } j \in B, v \in W \times X \times Y \tag{96}$$

We justify line 96 by noting that $\mathbb{U}^{\mathsf{V}_i|\mathsf{HX}}$ is a Markov kernel $H \times X \nrightarrow W \times X \times Y$, which is the same type as the map $\mathbf{Q} := h, a \mapsto \mathbb{P}_{\mathsf{X}=a}$, and in addition Equation 94 ensures that defining $\mathbb{U}^{\mathsf{V}_i|\mathsf{HX}} := \mathbf{Q}$ is consistent via Lemma 2.8.

Note that the assumptions of mutual independence $\mathsf{V}_i \perp\!\!\!\perp \mathsf{V}_{A\cup B\setminus\{i\}}|\mathsf{H}$ for $i \in A$ and $\mathsf{V}_j \perp\!\!\!\perp \mathsf{V}_{A\cup B\setminus\{j\}}|\mathsf{HX}_j$ for $j \in B$ are required for the existence of $\mathbb{U}^{\mathsf{V}_i|\mathsf{H}}$ and $\mathbb{U}^{\mathsf{V}_j|\mathsf{HX}_j}$ respectively.

Then Equation 94 becomes

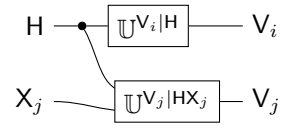
$$\mathbb{U}^{\mathsf{W}_j\mathsf{X}_j\mathsf{Y}_j|\mathsf{HX}_j}(w, x, y|h, a) = \mathbb{U}^{\mathsf{W}_i|\mathsf{H}}(w)\mathbb{U}^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{W}_i\mathsf{H}}(y|x, w, h)[\![x = a]\!] \quad i \in A, j \in B \tag{97}$$

The only difference here is that the coupling between distributions of observations and consequences via H is explicit.

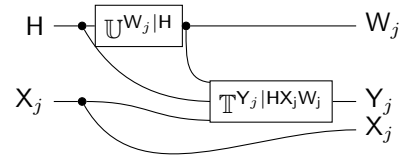
In most situations, A will be disjoint from B. While we don't necessarily want to rule out considering consequences to be equal to observations, we usually want to consider consequences that may take different values from observations.

For $i \in A$, $j \in B$, we can write $\mathbb{U}^{\mathsf{V}_i\mathsf{V}_j|\mathsf{H}\mathsf{X}_j}$ as follows

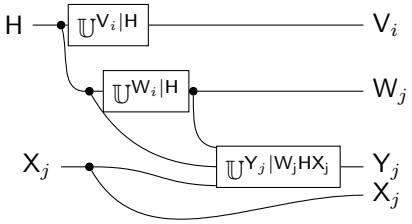
$$\mathbb{U}^{\mathsf{V}_i\mathsf{V}_j|\mathsf{H}\mathsf{X}_j} = \tag{98}$$



$$\mathbb{U}^{\mathsf{V}_j|\mathsf{H}\mathsf{X}_j} = \tag{99}$$

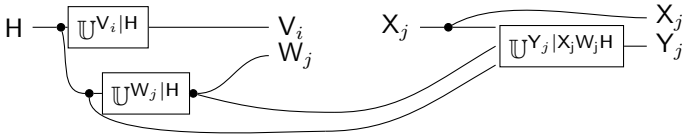


$$\implies \mathbb{U}^{\mathsf{V}_A\mathsf{V}_j|\mathsf{H}\mathsf{X}_j} = \tag{100}$$



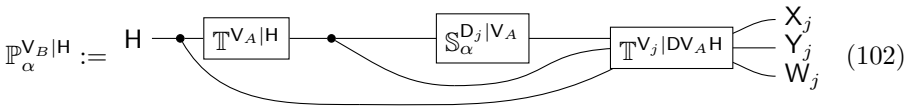
Note that we replace the single observation V_i with the full observations V_A as we will make use of them subsequently, and we can do this without issue due to the assumption of conditional independence among the V_ks. It will be sufficient to consider a single consequence V_j. Equation 100 defines a model $\mathbb{U}^{\cdot|\mathsf{H}\mathsf{X}_j}$ which relates observations to consequences in the manner suggested by Equation 94. We will call $\mathbb{U}$ a "CBN model". We note that the model in Equation 100 looks like a 2-comb:

$$\tag{101}$$

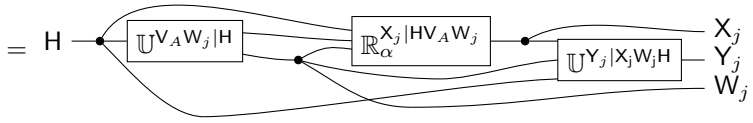


However, we have not at this point assumed that we have a convex set of strategies. Suppose we have some standard see-do model $\mathcal{M} := \{\mathbb{T}^{\mathsf{O}\mathsf{V}_B|\mathsf{H}\mathsf{D}}, \{\mathbb{S}_\alpha^{\mathsf{D}|\mathsf{O}} | \alpha \in \mathcal{A}\}\}$. The question we want to ask is: when can we posit a see-do model $\{\mathbb{U}^{\mathsf{V}_A\mathsf{V}_j|\mathsf{H}\mathsf{X}_j}, \{\mathbb{R}_\alpha^{\mathsf{X}_j|\mathsf{V}_A\mathsf{W}_j\mathsf{H}} | \alpha \in \mathcal{A}\}\}$ consistent with $\mathcal{M}$ in the sense that, for all

$\alpha \in \mathcal{A}$:



(102)



(103)

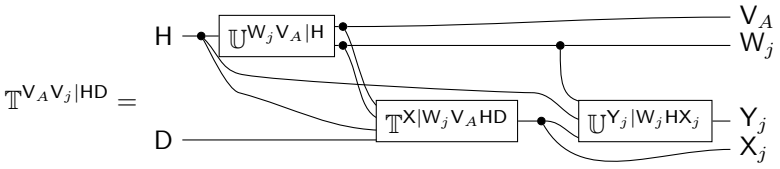
$$=: \mathbb{Q}_\alpha^{\mathsf{V}_B|\mathsf{H}} \tag{104}$$

I think reusing the same H between U and T is a mistake here. Maybe not a big problem, but ideally one would check!

Theorem 4.1. *Given a standard see-do model* $\mathcal{M} := \{\mathbb{T}^{\mathsf{O}\mathsf{V}_B|\mathsf{H}\mathsf{D}}, \{\mathbb{S}_\alpha^{\mathsf{D}|\mathsf{V}_A} | \alpha \in \mathcal{A}\}\}$ *and a CBN model* $\mathbb{U}^{\mathsf{V}_A\mathsf{V}_j|\mathsf{H}\mathsf{X}_j}$ *as defined in Equation 100, assuming* W_j *is a pre-choice variable, then there exists a see-do model* $\{\mathbb{U}^{\mathsf{V}_i\mathsf{V}_j|\mathsf{H}\mathsf{X}_j}, \{\mathbb{R}_\alpha^{\mathsf{X}_j|\mathsf{V}_A\mathsf{W}_j\mathsf{H}} | \alpha \in \mathcal{A}\}\}$ *consistent with* $\mathcal{M}$ *if and only if*

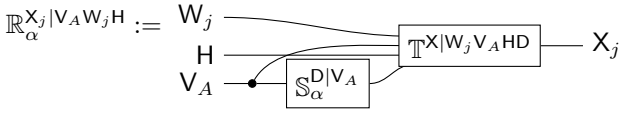
1. W_j *is a pre-choice variable, i.e.* $(\mathsf{V}_A, \mathsf{W}_j) \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$
2. $\mathbb{T}^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}} = \mathbb{U}^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}}$
3. $\mathsf{Y}_j \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j$
4. $\mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j} = \mathbb{U}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j}$

Proof. **If:** If all assumptions hold, we can write



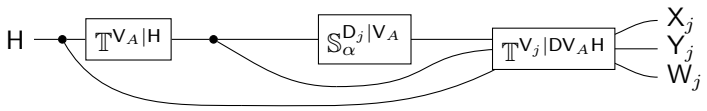
(105)

For each $\mathbb{S}_\alpha^{\mathsf{D}|\mathsf{V}_A}$, define

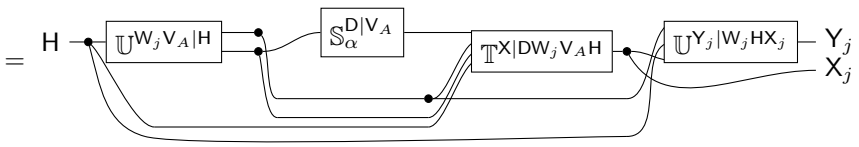


(106)

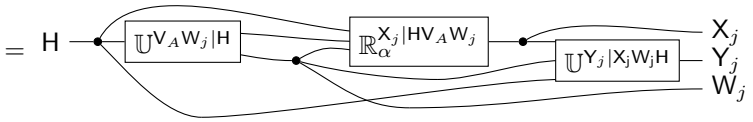
Then



(107)



(108)



(109)

Only if: Suppose assumption 1 does not hold. Then there exists some $d, d' \in D$, $w \in W$, $h \in H$ such that $\mathbb{T}^{\mathsf{W}_j|\mathsf{HD}}(w_j|h,d) \neq \mathbb{T}^{\mathsf{W}_j|\mathsf{HD}}(w|h,d')$. Then choose $\mathbb{S}_d^{\mathsf{D}|\mathsf{V}_A} : v_A \mapsto \delta_d$ and $\mathbb{S}_{d'}^{\mathsf{D}|\mathsf{V}_A} : v \mapsto \delta_{d'}$ for all $v \in V^{|A|}$. Then define

$$\mathbb{P}_d^{\mathsf{W}_j|\mathsf{H}}(w|h) = \mathbb{T}^{\mathsf{W}_j|\mathsf{HD}}(w_j|h,d) \tag{110}$$

$$\neq \mathbb{T}^{\mathsf{W}_j|\mathsf{HD}}(w_j|h,d') \tag{111}$$

$$= \mathbb{P}_{d'}^{\mathsf{W}_j|\mathsf{H}}(w|h) \tag{112}$$

But for any α, α', $\mathbb{Q}_\alpha^{\mathsf{W}_j|\mathsf{H}} = \mathbb{Q}_{\alpha'}^{\mathsf{W}_j|\mathsf{H}}$ as $\mathsf{W}_j \perp\!\!\!\perp_{\mathsf{U}} \mathsf{X}_j|\mathsf{H}$, so $\mathbb{Q} \neq \mathbb{P}$. Suppose assumption 1 holds but assumption 2 does not. Then for any α

$$\mathbb{P}_\alpha^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}} = \mathbb{T}^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}} \tag{113}$$

$$\neq \mathbb{U}^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}} \tag{114}$$

$$= \mathbb{Q}_\alpha^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}} \tag{115}$$

Suppose assumption 3 does not hold. Then there is some $d, d' \in D$, $w \in W$, $h \in H$, $v \in V^{|A|}$, $x \in X$ and $y \in Y$ such that

$$\mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{HX}_j\mathsf{D}}(y|w,v,h,x,d) \neq \mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{HX}_j\mathsf{D}}(y|w,v,h,x,d') \tag{116}$$

$$\text{and } \mathbb{T}^{\mathsf{X}_j\mathsf{W}_j\mathsf{V}_A|\mathsf{HD}}(x,w,v|h,d) > 0 \tag{117}$$

$$\text{and } \mathbb{T}^{\mathsf{X}_j\mathsf{W}_j\mathsf{V}_A|\mathsf{HD}}(x,w,v|h,d') > 0 \tag{118}$$

$$\tag{119}$$

The latter conditions hold as if Equation 116 only held on sets of measure 0 then we could choose versions of the conditional probabilities such that the independence held.

Then

$$\mathbb{P}_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) = \mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x,d) \tag{120}$$

$$\neq \mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x,d') \tag{121}$$

$$= \mathbb{P}_{d'}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) \tag{122}$$

$$\implies \mathbb{P}_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) \neq \mathbb{Q}_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) \tag{123}$$

$$\text{or } \mathbb{P}_{d'}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) \neq \mathbb{Q}_{d'}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) \tag{124}$$

As the conditional probabilities disagree on a positive measure set, $\mathbb{P} \neq \mathbb{Q}$.

Suppose assumption 3 holds but assumption 4 does not. Then for some $h \in H$, some $w \in W$, $v \in V^{|A|}$, $x \in X$ with positive measure and some $y \in Y$

$$\mathbb{P}_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) = \mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j}(y|w,v,h,x) \tag{125}$$

$$\neq \mathbb{U}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j}(y|w,v,h,x) \tag{126}$$

$$\neq modelQ_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) \tag{127}$$

□

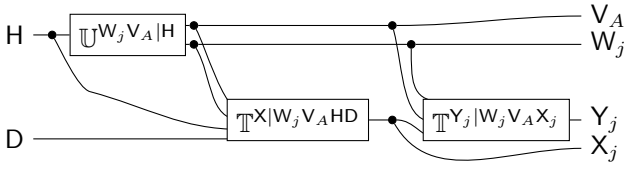
Conditional independences like $(\mathsf{V}_A, \mathsf{W}_j) \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$ and $\mathsf{Y}_j \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j$ bear some resemblance to the condition of "limited unresponsiveness" proposed by Heckerman and Shachter (1995). They are conceptually similar in that they indicate that a particular variable does not "depend on" a decision D in some sense. As Heckerman points out, however, limited unresponsiveness is not equivalent to conditional independence. We tentatively speculate that there may be a relation between our "pre-choice variables" $(\mathsf{W}_j, \mathsf{V}_A, \mathsf{H})$ and the "state" in Heckerman's work crucial for defining limited unresponsiveness.

4.1 Proxy control

We say that $(\mathsf{V}_A, \mathsf{W}_j) \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$ expresses the notion that W_j is a *pre-choice variable* and $(\mathsf{W}_j, \mathsf{V}_A, \mathsf{X}_j)$ are *proxies for* D with respect to Y under conditions of full information. To justify this terminology, we note that under a strong assumption of identifiability $\mathsf{Y}_j \perp\!\!\!\perp \mathsf{H}|\mathsf{W}_j\mathsf{V}_A\mathsf{X}_j$ (i.e. the observed data allow us

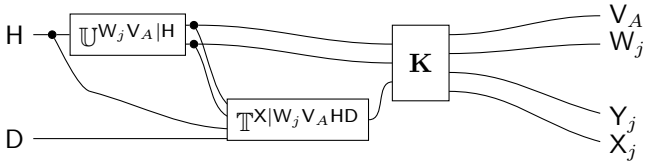
to identify H for the purposes of determining $\mathsf{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{X}_j\mathsf{H}}$), then we can write

$$\mathbb{T}^{\mathsf{V}_A\mathsf{V}_B|\mathsf{HD}} =$$



(128)

$$=$$



$$= \mathbb{T}^{\mathsf{V}_A\mathsf{W}_j\mathsf{X}_j|\mathsf{HD}}\mathbf{M}$$

(129)

That is, under conditions of full information, knowing how to control the proxies $(\mathsf{W}_j, \mathsf{V}_A, \mathsf{X}_j)$ is sufficient to control Y. This echoes Pearl (2018)'s view on causal effects representing "stable characteristics":

> Smoking cannot be stopped by any legal or educational means available to us today; cigarette advertising can. That does not stop researchers from aiming to estimate "the effect of smoking on cancer," and doing so from experiments in which they vary the instrumentcigarette advertisementnot smoking. The reason they would be interested in the atomic intervention $P(\text{cancer}|do(\text{smoking}))$ rather than (or in addition to) $P(\text{cancer}|do(\text{advertising}))$ is that the former represents a stable biological characteristic of the population, uncontaminated by social factors that affect susceptibility to advertisement, thus rendering it transportable across cultures and environments. With the help of this stable characteristic, one can assess the effects of a wide variety of practical policies, each employing a different smoking-reduction instrument.

5 Potential outcomes

Like causal Bayesian networks, causal models in the potential outcomes framework typically do not include any variables representing what we call "consequences". A potential outcomes model features a sequence of observable variables $(\mathsf{Y}_i, \mathsf{X}_i, \mathsf{Z}_i)_{i\in[n]}$ and a collection of potential outcomes $(\mathsf{Y}_i^x)_{x\in X, i\in[n]}$. Also like causal Bayeisan networks, we think that introducing the idea of consequences clarifies the meaning of potential outcomes models.

We begin with a formal definition of potential outcomes, but as we will discuss this formal definition is not enough on its own to tell us what potential outcomes are. Formally, potential outcomes of Y taking values in Y with respect to X taking values in X are a variable Y^X taking values in Y^X such that Y is related to Y^X and X via a *selector*.

Definition 5.1 (Selector)**.** Given variables $\mathsf{X} : \Omega \to X$ and $\{\mathsf{Y}^x : \Omega \to Y | x \in X\}$, define $\mathsf{Y}^X : (\mathsf{Y}^x)_{x \in X}$. The selector $\pi : X \times Y^X \to Y$ is the function that sends $(x, y^1, ..., y^{|X|}) \to y^x$.

Definition 5.2 (Potential outcomes: formal requirement)**.** Given variables $\mathsf{Y} : \Omega \to Y$ and $\mathsf{X} : \Omega \to X$, we introduce a collection of latent variables called *potential outcomes* $\mathsf{Y}^X := (\mathsf{Y}^x)_{x \in X}$ such that $\mathsf{Y} = \pi \circ (\mathsf{X}, \mathsf{Y}^X)$. A *potential outcomes model* is any consistent model of Y, X and Y^X.

Lemma 5.3 shows we can always define trivial potential outcomes of Y with respect to X by taking the product of $|X|$ copies of Y. We need some other constraint on the values of potential outcomes besides the formal definition 5.2 if we want them to be informative.

Lemma 5.3 (Trivial formal potential outcomes)**.** *For any variables $\mathsf{Y} : \Omega \to Y$, $\mathsf{X} : \Omega \to X$ and $\mathsf{W} : \Omega \to W$, we can always define potential outcomes Y_X such that any consistent model $\mathbb{K}^{\mathsf{YX}|\mathsf{W}}$ can be extended to a consistent model of $\mathbb{K}^{\mathsf{YXY}^X|\mathsf{W}}$.*

Proof. Define $\mathsf{Y}^X := (\mathsf{Y})_{x \in X}$. Then we can consistently extend $\mathbb{K}^{\mathsf{YX}|\mathsf{W}}$ to $\mathbb{K}^{\mathsf{YXY}^X|\mathsf{W}}$ by repeated application of Lemma 2.7. □

The trivial potential outcomes of Lemma 5.3 are in many cases unsatisfactory for what we want potential outcomes to represent. Thus Definition 5.2 is incomplete. In common with observable variables, the definition of potential outcomes involves both the formal requirement of Definition 5.2, and an indication of the parts of the real world that they model. Unlike observable variables, the "part of the world" that potential outcomes model will not at any point resolve to a canonical value. We say the potential outcome $\mathsf{Y}^x := \pi(x, \mathsf{Y})$ is "the value that Y would take if X were x, whether or not X actually takes the value x". We will call this additional element of the definition of potential outcomes the *counterfactual extension*.

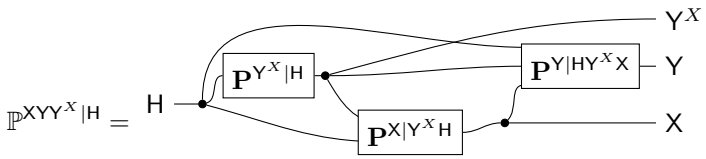
Definition 5.4 (What potential outcomes model: counterfactual extension)**.** Given observables X, Y and Y^X, Y^X are potential outcomes if they satisfy Definition 5.2 and for all $x \in X$, the individual potential outcome $\mathsf{Y}^x := \pi(x, \mathsf{Y})$ models the value Y would take if X took the value x.

Because observables resolve to a single canonical value, the conditional in Definition 5.4 is eventually satisfied for exactly one $x \in X$, at which point $\mathsf{Y}^{x'}$ for all $x' \neq x$ are guaranteed not to resolve. Nevertheless, we can maybe draw some conclusions about Y^X from Definition 5.4. For example, it seems unreasonable in light of this definition to assert that Y^x is *necessarily* identical to Y for all $x \in X$, which rules out the strictly trivial potential outcomes of Lemma 5.3.

We will note at this point that if X refers to a person's body mass index and Y to an indicator of whether or not they experience heart disease, it is metaphysically subtle to say whether Y^X is well-defined with regard to Definitions 5.2 and 5.4 together. Recall that there are multiple ways that a given level of body

mass index (X) could be achieved. One might say that, when there are multiple possibile paths, there is no unique way to choose a path. However, a very similar argument can be made that whenever there are multiple possible values of Y^x (which is whenever X does not take the value x), then there is no unique choice of Y^x, which implies that the full set of potential outcomes Y^X is *almost never well-defined*. Alternatively, if there is some method of making a canonical choice of Y^x, then perhaps this same method can also make a canonical choice of which path was taken to achieve this value of X.

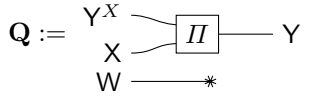
We will set Definition 5.4 aside and propose an alternative decision-theoretic extension of the definition of potential outcomes. To motivate this proposal, we first note that, if we are using potential outcomes Y^X to model an observation of X and Y only conditional on some hypothesis (or parameter) H, then by repeated application of Lemma 2.19, we can represent the model $\mathbb{P}^{\mathsf{XYY}^X|\mathsf{H}}$ of these variables as



(130)

For any collection of representative kernels $\mathbf{T}^{\mathsf{Y}^X|\mathsf{H}}$, $\mathbf{T}^{\mathsf{X}|\mathsf{Y}^X\mathsf{H}}$ and $\mathbf{T}^{\mathsf{Y}|\mathsf{HY}^X\mathsf{X}}$. We can simplify Equation 130 somewhat. Firstly, $\mathbb{P}^{\mathsf{Y}|\mathsf{HY}^X\mathsf{X}}$ must always be represented a *selector kernel* $\mathit{\Pi} : X \times Y^{|X|} \twoheadrightarrow Y$, as shown by Lemma 5.5.

Lemma 5.5 (Selector kernel)**.** *Let the selector kernel* $\mathit{\Pi} : X \times Y^X \twoheadrightarrow Y$ *be defined by* $\mathit{\Pi}^y_{(x,y^X)} = [\![\pi(x, y^X) = y]\!]$*. Given* X*,* Y*, potential outcomes* Y^X *and arbitrary* W*, defining* $\mathbf{Q} : X \times Y^X \times W \twoheadrightarrow Y$ *by*



(131)

$$\iff \tag{132}$$

$$\mathbf{Q}^y_{(y^X,x,w)} = \mathit{\Pi}^y_{(x,y^X)} \qquad \forall y, y^X, x, w \tag{133}$$

Then any potential outcomes model $\mathbb{T}^{\mathsf{YY}^X\mathsf{X}|\mathsf{W}}$ *must have the property that, for all* x, w, y^X *and* y*,* $\mathbf{Q}$ *is a representative of* $\mathbb{T}^{\mathsf{Y}|\mathsf{Y}^X\mathsf{XW}}$*.*

Proof. Recall $\mathsf{Y} = \pi \circ (\mathsf{X}, \mathsf{Y}^X)$. Thus consistency implies that $\mathsf{Y} \overset{a.s.}{=} \pi \circ (\mathsf{X}, \mathsf{Y}^X)$ for all $(x, y^X, w) \in \text{Range}(\mathsf{X}) \times \text{Range}(\mathsf{Y}) \times \text{Range}(\mathsf{W})$ such that $\mathsf{X}^{-1}(x) \cap (\mathsf{Y}^X)^{-1}(y^X) \cap \mathsf{W}^{-1}(w) \neq \emptyset$. However, wherever $\mathsf{X}^{-1}(x) \cap (\mathsf{Y}^X)^{-1}(y^X) \cap \mathsf{W}^{-1}(w) = \emptyset$, consistency implies $\mathbb{T}^{\mathsf{YY}^X\mathsf{X}|\mathsf{W}}(y, y^X, x|w) = 0$ and so $\mathbb{T}^{\mathsf{Y}|\mathsf{Y}^X\mathsf{XW}}$ is arbitrary on this collection of values. Equations 131 and 133 are equivalent to the statement $\mathsf{Y} \overset{a.s.}{=} \pi \circ (\mathsf{X}, \mathsf{Y}^X)$. □

Thus we can without loss of generality choose Π to represent $\mathbb{T}^{\mathsf{Y}|\mathsf{Y}^X\mathsf{XW}}$. We observe that when Rubin (2005) describes a potential outcomes model, he calls $\mathbb{T}^{\mathsf{Y}^X|\mathsf{H}}$ "the science" and $\mathbb{T}^{\mathsf{X}|\mathsf{HY}^X}$ the "selection function". He goes on to explain that the science "is not affected by how or whether we try to learn about it".

We propose a definition of potential outcomes that enshrines the stability of "the science".

Definition 5.6. Potential outcomes: decision theoretic extension Given a standard decision problem $\{\mathbb{T}^{\mathsf{WZ}|\mathsf{HD}}, \{\mathbb{S}_\alpha\}_{\alpha\in A}\}$, Y^X is a potential outcome for Y with respect to X if it satisfies Definition 5.2 and is a prechoice variable; that is, $(\mathsf{Y}^X, \mathsf{W}) \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$.

Owing to the subtlety of interpreting Definition 5.4, we don't know a straightforward argument to the effect that Definition 5.6 is implied by it. Besides the fact that it seems to formalise the idea that the distribution of potential outcomes is unaffected by our actions, we will point out that a key feature of prechoice variables – decisions can be chosen so that they are random with respect to all prechoice variables – is used in practice to justify the assumption of ignorability in randomised experiments.

Definition 5.6 can sometimes (but not always) rule out potential outcomes if there is more than one way to achieve a given value of X. Recall that Hernán and Taubman (2008) argued potential outcomes are "ill-defined" in the presence of multiple treatments.

Example 5.7. Suppose we have a standard decision problem $\{\mathbb{T}^{\mathsf{WZ}|\mathsf{HD}}, \{\mathbb{S}_\alpha\}_{\alpha\in A}\}$ where observations are W, consequences Z, hypotheses H and decisions $\mathsf{D} \in \{0,1,2,3\}$. Suppose we also have some $\mathsf{X} \in \{0,1\}$, Y such that $\mathbb{T}^{\mathsf{X}|\mathsf{HWD}}(x|h,w,d) = [\![x = d \bmod 2]\!]$ for all h, w and, for some y

$$\mathbb{T}^{\mathsf{Y}|\mathsf{HWXD}}(y|h,w,0,0) \neq \mathbb{T}^{\mathsf{Y}|\mathsf{HWXD}}(y|h,w,0,2) \tag{134}$$

Then we can consider strategies $\mathbb{S}_0^{\mathsf{D}|\mathsf{W}} := w \mapsto \delta_0$ and $\mathbb{S}_2^{\mathsf{D}|\mathsf{W}} := w \mapsto \delta_2$. By assumption,

$$\mathbb{P}_0^{\mathsf{Y}|\mathsf{HD}}(y|h,0) = \sum_{x\in\{0,1\}, w\in W} \mathbb{T}^{\mathsf{W}|\mathsf{H}}(w|h)\mathbb{S}_0^{\mathsf{D}|\mathsf{W}}(0|w)\mathbb{T}^{\mathsf{X}|\mathsf{HWD}}(x|h,w,0)\mathbb{T}^{\mathsf{Y}|\mathsf{HWXD}}(y|h,w,x,0) \tag{135}$$

$$= \mathbb{T}^{\mathsf{Y}|\mathsf{HWXD}}(y|h,w,0,0) \tag{136}$$

$$\neq \mathbb{P}_2^{\mathsf{Y}|\mathsf{HD}} \tag{137}$$

Suppose we had some potential outcomes Y^X for Y with respect to X. Then, by

assumption

$$\mathbb{P}_0^{\mathsf{Y}|\mathsf{HD}}(y|h,0) = \sum_{y^X \in Y^2, x \in \{0,1\}} \mathbb{T}^{\mathsf{Y}^X|\mathsf{H}}(y^X|h)\mathbb{T}^{\mathsf{X}|\mathsf{HDY}^X}(x|h,0,y^X)\Pi(y|x,y^X) \tag{138}$$

$$= \sum_{y^X} \mathbb{T}^{\mathsf{Y}^X|\mathsf{H}}(y^X|h)\Pi(y|0,y^X) \tag{139}$$

$$= \sum_{y^X \in Y^2, x \in \{0,1\}} \mathbb{T}^{\mathsf{Y}^X|\mathsf{H}}(y^X|h)\mathbb{T}^{\mathsf{X}|\mathsf{HDY}^X}(x|h,2,y^X)\Pi(y|x,y^X) \tag{140}$$

$$= \mathbb{P}_2^{\mathsf{Y}|\mathsf{HD}} \tag{141}$$

Here we use the property $\mathsf{Y}^X \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$, implied by the assumption that Y^X is a prechoice variable. Equations 137 and 141 are clearly contradictory, thus there can be no potential outcomes Y^X in this example.

I think I asked the wrong question here – should've asked when I can extend a see-do model with additonal pre-choice variables. I think it's possible to always choose some deterministic potential outcomes.

Theorem 5.8 (Existence of potential outcomes)**.** *Suppose we have a standard decision problem* $\{\mathbb{T}^{\mathsf{WZ}|\mathsf{HD}}, \{\mathbb{S}_\alpha\}_{\alpha\in A}\}$*, and let* U *be the sequence of all prechoice variables. For some* Y *and* X*, there exist potential outcomes* Y^X *in the sense of Definition 5.6 if and only if* $\mathbb{T}^{\mathsf{Y}|\mathsf{UX}}$ *exists and is deterministic.*

Proof. If: If $\mathbb{T}^{\mathsf{Y}|\mathsf{UX}}$ exists and is deterministic then there exists some $f : U \times X \to Y$ such that $\mathsf{Y} \overset{a.s.}{=} f \circ (\mathsf{U},\mathsf{X})$. Let $\mathsf{Y}^X := (f(\mathsf{U},x))_{x\in X}$. Then $\pi \circ (\mathsf{X},\mathsf{Y}^X) = f(\mathsf{U},\mathsf{X}) \overset{a.s.}{=} \mathsf{Y}$.

Only if: By definition, $\mathsf{Y}^X = g \circ \mathsf{U}$. From Lemma 5.5, $\mathbb{T}^{\mathsf{Y}|\mathsf{XY}^X}$ exists and is deterministic. Thus $\mathbb{T}^{\mathsf{Y}|\mathsf{XW}}$ also exists and is also deterministic. □

Corollary 5.9. *Potential outcomes* Y^X *in the sense of Definition 5.6 exist only if*

$$\mathsf{Y} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{WX} \tag{142}$$

Proof. $\mathbb{T}^{\mathsf{Y}|\mathsf{UX}}$ exists only if $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{UX}$. □

Note the similarity between Equation 142 and the condition for proxy control in the previous section. Indeed, the two are identical if we identify U with $(\mathsf{W}_j, \mathsf{V}_A, \mathsf{X}_j)$.

6 Appendix:see-do model representation

Update notation

Theorem 6.1 (See-do model representation)**.** *Suppose we have a decision problem that provides us with an observation $x \in X$, and in response to this we can select any decision or stochastic mixture of decisions from a set D; that is we can choose a "strategy" as any Markov kernel $\mathbf{S} : X \to \Delta(D)$. We have a utility function $u : Y \to \mathbb{R}$ that models preferences over the potential consequences of our choice. Furthermore, suppose that we maintain a denumerable set of hypotheses H, and under each hypothesis $h \in H$ we model the result of choosing some strategy $\mathbf{S}$ as a joint probability over observations, decisions and consequences $\mathbb{P}_{h,\mathbf{S}} \in \Delta(X \times D \times Y)$.*

Define X, Y and D such that $\mathsf{X}_{xdy} = x$, $\mathsf{Y}_{xdy} = y$ and $\mathsf{D}_{xdy} = d$. Then making the following additional assumptions:

1. *Holding the hypothesis h fixed the observations as have the same distribution under any strategy: $\mathbb{P}_{h,\mathbf{S}}[\mathsf{X}] = \mathbb{P}_{h,\mathbf{S}''}[\mathsf{X}]$ for all $h, \mathbf{S}, \mathbf{S}'$ (observations are given "before" our strategy has any effect)*
2. *The chosen strategy is a version of the conditional probability of decisions given observations: $\mathbf{S} = \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]$*
3. *There exists some strategy $\mathbf{S}$ that is strictly positive*
4. *For any $h \in H$ and any two strategies $\mathbf{Q}$ and $\mathbf{S}$, we can find versions of each disintegration such that $\mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{DX}] = \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{DX}]$ (our strategy tells us nothing about the consequences that we don't already know from the observations and decisions)*

Then there exists a unique see-do model $(\mathbf{T}, \mathsf{H}', \mathsf{D}', \mathsf{X}', \mathsf{Y}')$ such that $\mathbb{P}_{h,\mathbf{S}}[\mathsf{XDY}]^{ijk} = \mathbf{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbf{S}^j_i \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ijk}$.

Proof. Consider some probability $\mathbb{P} \in \Delta(X \times D \times Y)$. By the definition of disintegration (section **??**), we can write

$$\mathbb{P}[\mathsf{XDY}]^{ijk} = \mathbb{P}[\mathsf{X}]^i \mathbb{P}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{P}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \tag{143}$$

Fix some $h \in H$ and some strictly positive strategy $\mathbf{S}$ and define $\mathbf{T} : H \times D \to \Delta(X \times Y)$ by

$$\mathbf{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^l_{kj} \tag{144}$$

Note that because $\mathbf{S}$ is strictly positive and by assumption $\mathbf{S} = \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]$, $\mathbb{P}_{h,\mathbf{S}}[\mathsf{D}]$ is also strictly positive. Therefore $\mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{D}]$ is unique and therefore $\mathbf{T}$ is also unique.

Define X' and Y' by $\mathsf{X}'_{xy} = x$ and $\mathsf{Y}'_{xy} = y$. Define H' and D' by $\mathsf{H}'_{hd} = h$ and $\mathsf{D}'_{hd} = d$.

We then have

$$\mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj} = \mathbf{T}\underline{\mathsf{X}}'^k_{hj} \tag{145}$$
$$= \sum_l \mathbf{T}^{kl}_{hj} \tag{146}$$
$$= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^k \tag{147}$$
$$= \mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj'} \tag{148}$$

Thus $\mathsf{X}' \perp\!\!\!\perp_{\mathbf{T}} \mathsf{D}'|\mathsf{H}'$ and so $\mathbf{T}[\mathsf{X}'|\mathsf{H}']$ exists (section 2.10) and $(\mathbf{T}, \mathsf{H}', \mathsf{D}', \mathsf{X}', \mathsf{Y}')$ is a see-do model.

Applying Equation 143 to $\mathbb{P}_{h,\mathbf{S}}$:

$$\mathbb{P}_{h,\mathbf{S}}[\mathsf{XDY}]^{ijk} = \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \tag{149}$$
$$= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \tag{150}$$
$$= \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]^j_i \mathbf{T}[\mathsf{X}'\mathsf{Y}'|\mathsf{H}'\mathsf{D}']^{ik}_{hj} \tag{151}$$
$$= \mathbf{S}^j_i \mathbf{T}[\mathsf{X}'\mathsf{Y}'|\mathsf{H}'\mathsf{D}']^{ik}_{hj} \tag{152}$$
$$= \mathbf{S}^j_i \mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^i_{hj} \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} \tag{153}$$
$$= \mathbf{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbf{S}^j_i \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} \tag{154}$$

Consider some arbitrary alternative strategy $\mathbf{Q}$. By assumption

$$\mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i = \mathbb{P}_{h,\mathbf{Q}}[\mathsf{X}]^i \tag{155}$$
$$\mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} = \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \text{ for some version of } \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}] \tag{156}$$

It follows that, for some version of $\mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}]$,

$$\mathbf{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbf{Q}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}]^l_{kj} \tag{157}$$

Then by substitution of $\mathbf{Q}$ for $\mathbf{S}$ in Equation 149 and working through the same steps

$$\mathbb{P}_{h,\mathbf{S}}[\mathsf{XDY}]^{ijk} = \mathbf{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbf{Q}^j_i \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} \tag{158}$$

As $\mathbf{Q}$ was arbitrary, this holds for all strategies. □

7 Appendix: Counterfactual representation

Definition 7.1 (Parallel potential outcomes)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, a collection of variables $\{\mathsf{Y}_i, \mathsf{Y}(W), \mathsf{W}_i\}$, $i \in [n]$, where Y_i and $\mathsf{Y}(W)$ are random variables and W_i could be either a state or random variables is a *parallel potential outcome submodel* if $\mathbf{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]$ exists and $\mathbf{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]_{kj_1j_2\ldots j_{|W|}} = \delta[j_k]$.

How this will change: a parallel potential outcomes model is a comb $\mathbb{K}[\mathsf{Y}(W)|\mathsf{H}] \rightrightarrows \mathbb{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]$.

A parallel potential outcomes model features a sequence of n "parallel" outcome variables Y_i and n "regime proposals" W_i, with the property that if the regime proposal $\mathsf{W}_i = w_i$ then the corresponding outcome $\mathsf{Y}_i \overset{a.s.}{=} \mathsf{Y}(w_i)$. We can identify a particular index, say $n = 1$, with the actual world and the rest of the indices with supposed worlds. Thus Y_1 represents the value of TYT in the actual world and Y_i $i \neq 1$ represents TYT under a supposed regime W_i. Given such an interpretation, the fact that $\mathsf{Y}_i \overset{a.s.}{=} \mathsf{Y}(w_i)$ can be interpreted as assuming "for all w, if the supposed regime W_i is w then the corresponding outcome will be almost surely equal to $\mathsf{Y}(w)$, regardless of the value of the actual regime W_1", which is our original counterfactual assumption.

We do not intend to defend this as the only way that counterfactuals can be modeled, or even that it is appropriate to capture the idea of counterfactuals at all. It is simply a way that we can model the counterfactual assumption typically associated with potential outcomes. We will show show that parallel potential outcome submodels correspond precisely to *extendably exchangeable* and *deterministically reproducible* submodels of Markov kernel spaces.

7.1 Parallel potential outcomes representation theorem

Exchangeble sequences of random variables are sequences whose joint distribution is unchanged by permutation. Independent and identically distributed random variables are one example: if X_1 is the result of the first flip of a coin that we know to be fair and X_2 is the second flip then $\mathbb{P}[\mathsf{X}_1\mathsf{X}_2] = \mathbb{P}[\mathsf{X}_2\mathsf{X}_1]$. There are also many examples of exchangeable sequences that are not mutually independent and identically distributed – for example, if we want to use random variables Y_1 and Y_2 to model our subjective uncertainty regarding two flips of a coin of unknown fairness, we regard our initial uncertainty for each flip to be equal $\mathbb{P}[\mathsf{Y}_1] = \mathbb{P}[\mathsf{Y}_2]$ and we our state of knowledge of the second flip after observing only the first will be the same as our state of knowledge of the first flip after observing only the second $\mathbb{P}[\mathsf{Y}_2|\mathsf{Y}_1] = \mathbb{P}[\mathsf{Y}_1|\mathsf{Y}_2]$, then our model of subjective uncertainty is exchangeable.

De Finetti's representation theorem establishes the fact that any infinite exchangeable sequence $\mathsf{Y}_1, \mathsf{Y}_2, ...$ can be modeled by the product of a *prior* probability $\mathbb{P}[\mathsf{J}]$ with J taking values in the set of marginal probabilities $\Delta(Y)$ and a conditionally independent and identically distributed Markov kernel $\mathbb{P}[\mathsf{Y}_A|\mathsf{J}]_j^{y_A} = \prod_{i \in A} \mathbb{P}[\mathsf{Y}_1|\mathsf{J}]_j^{y_i}$.

We extend the idea of exchangeable sequences to cover both random variables and state variables, and we show that a similar representation theorem holds for potential outcomes. De Finetti's original theorem introduced the variable J that took values in the set of marginal distributions over a single observation; the set of potential outcome variables plays an analagous role taking values in the set of functions from propositions to outcomes.

The representation theorem for potential outcomes is somewhat simpler that

De Finetti's original theorem due to the fact that potential outcomes are usually assumed to be *deterministically reproducible*; in the parallel potential outcomes model, this means that for $j \neq i$, if W_j and W_i are equal then Y_j and Y_i will be almost surely equal. This assumption of determinism means that we can avoid appeal to a law of large numbers in the proof of our theorem.

An interesting question is whether there is a similar representation theorem for potential outcomes without the assumption of deterministic reproducibility. I'm reasonably confident that this is a straightforward corollary of the representation theorem proved in my thesis. However, this requires maths not introduced in this draft of the paper.

Extendably exchangeable sequences can be permuted without changing their conditional probabilities, and can be extended to arbitrarily long sequences while maintaining this property. We consider here sequences that are exchangeable conditional on some variable; this corresponds to regular exchageability if the conditioning variable is $*$ where $*_i = 1$.

Definition 7.2 (Exchangeability)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ with Y_i random variables is *exchangeable* conditional on Z if, defining $\mathsf{Y}_{[n]} = (\mathsf{Y}_i)_{i\in[n]}$ and $\mathsf{D}_{[n]} = (\mathsf{D}_i)_{i\in[n]}$, $\mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$ exists and for any bijection $\pi : [n] \to [n]$ $\mathbf{K}[\mathsf{Y}_{\pi([n])}|\mathsf{D}_{\pi([n])}\mathsf{Z}] = \mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$.

Definition 7.3 (Extension)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, $(\mathbf{K}', E', F')$ is an *extension* of $(\mathbf{K}, E, F)$ if there is some random variable X and some state variable U such that $\mathbf{K}'[\mathsf{X}|\mathsf{U}]$ exists and $\mathbf{K}'[\mathsf{X}|\mathsf{U}] = \mathbf{K}$.

If $(\mathbf{K}', E', F')$ is an extension of $(\mathbf{K}, E, F)$ we can identify any random variable Y on $(\mathbf{K}, E, F)$ with $\mathsf{Y} \circ \mathsf{X}$ on $(\mathbf{K}', E', F')$ and any state variable D with $\mathsf{D}\circ\mathsf{U}$ on $(\mathbf{K}', E', F')$ and under this identification $\mathbf{K}'[\mathsf{Y}\circ\mathsf{X}|\mathsf{D}\circ\mathsf{E}]$ exists iff $\mathbf{K}[\mathsf{Y}|\mathsf{D}]$ exists and $\mathbf{K}'[\mathsf{Y} \circ \mathsf{X}|\mathsf{D} \circ \mathsf{E}] = \mathbf{K}[\mathsf{Y}|\mathsf{D}]$. To avoid proliferation of notation, if we propose $(\mathbf{K}, E, F)$ and later an extension $(\mathbf{K}', E', F')$, we will redefine $\mathbf{K} := \mathbf{K}'$ and $\mathsf{Y} := \mathsf{Y} \circ \mathsf{X}$ and $\mathsf{D} := \mathsf{D} \circ \mathsf{E}$.

I think this is a very standard thing to do – propose some X and $\mathbb{P}(\mathsf{X})$ then introduce some random variable Y and $\mathbb{P}(\mathsf{XY})$ as if the sample space contained both X and Y all along.

Definition 7.4 (Extendably exchangeable)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ and a state variable Z with Y_i random variables is *extendably exchangeable* if there exists an extension of $\mathbf{K}$ with respect to which $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in\mathbb{N}}$ is exchangeable conditional on Z.

Here that we identify Z and $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ defined on the extension with the original variables defined on $(\mathbf{K}, E, F)$ while $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in\mathbb{N}\setminus[n]}$ may be defined only on the extension.

Deterministically reproducible sequences have the property that repeating the same decision gets the same response with probability 1. This could be a model of an experiment that exhibits no variation in results (e.g. every time I

put green paint on the page, the page appears green), or an assumption about collections of "what-ifs" (e.g. if I went for a walk an hour ago, just as I actually did, then I definitely would have stubbed my toe, just like I actually did). Incidentally, many consider that this assumption is false concering what-if questions about things that exhibit quantum behaviour.

Definition 7.5 (Deterministically reproducible)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ with Y_i random variables is *deterministically reproducible* conditional on Z if $n \geq 2$, $\mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$ exists and $\mathbf{K}[\mathsf{Y}_{\{i,j\}}|\mathsf{D}_{\{i,j\}}\mathsf{Z}]^{lm}_{kk} = [\![l = m]\!]\mathbf{K}[\mathsf{Y}_i|\mathsf{D}_i\mathsf{Z}]^l_k$ for all i, j, k, l, m.

Theorem 7.6 (Potential outcomes representation)**.** *Given a Markov kernel space $(\mathbf{K}, E, F)$ along with a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ with $n \geq 2$ and a conditioning variable Z, $(\mathbf{K}, E, F)$ can be extended with a set of variables $\mathsf{Y}(D) := (\mathsf{Y}(i))_{i\in D}$ such that $\{\mathsf{Y}_i, \mathsf{Y}(D), \mathsf{D}_i\}$ is a parallel potential outcome submodel if and only if $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ is extendably exchangeable and deterministically reproducible conditional on Z.*

Proof. If: Because $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ is extendably exchangeable, we can without loss of generality assume $n \geq |D|$.

Let $e = (e_i)_{i\in[|D|]}$. Introduce the variable $\mathsf{Y}(i)$ for $i \in D$ such that $\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_{[D]}\mathsf{Z}]_{ez} = \mathbf{K}[\mathsf{Y}_D|\mathsf{D}_D\mathsf{Z}]_{ez}$ and introduce X_i, $i \in D$ such that $\mathbf{K}[\mathsf{X}_i|\mathsf{D}_i\mathsf{Z}\mathsf{Y}(D)]^{x_i}_{e_i z j_1 \dots j_{|D|}} = \delta[j_{e_i}]^{x_i}$. Clearly $\{\mathsf{X}_{[n]}, \mathsf{D}_{[n]}, \mathsf{Y}(D)\}$ is a parallel potential outcome submodel. We aim to show that $\mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}] = \mathbf{K}[\mathsf{X}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$.

Let $y := (y_i)_{i\in|D|} \in Y^{|D|}$, $d := (d_i)_{i\in[n]} \in D^{[n]}$, $x := (x_i)_{i\in[n]} \in Y^{[n]}$.

$$\mathbf{K}[\mathsf{X}_n|\mathsf{D}_n\mathsf{Z}]^x_{dz} = \sum_{y\in Y^{|D|}} \mathbf{K}[\mathsf{X}_{[n]}|\mathsf{D}_n\mathsf{Z}\mathsf{Y}(D)]^x_{dzy}\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_{[n]}\mathsf{Z}]^y_{dz} \tag{159}$$

$$= \sum_{y\in Y^{|D|}} \prod_{i\in[n]} \delta[y_{d_i}]^{x_i}\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_n\mathsf{Z}]^y_{dz} \tag{160}$$

Wherever $d_i = d_j := \alpha$, every term in the above expression will contain the product $\delta[\alpha]^{x_i}\delta[\alpha]^{x_j}$. If $x_i \neq x_j$, this will always be zero. By deterministic reproducibility, $d_i = d_j$ and $x_i \neq x_j$ implies $\mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]_d z^x = 0$ also. We need to check for equality for sequences x and d such that wherever $d_i = d_j$, $x_i = x_j$. In this case, $\delta[\alpha]^{x_i}\delta[\alpha]^{x_j} = \delta[\alpha]^{x_i}$. Let $Q_d \subset [n] := \{i| \nexists i \in [n] : j < i \ \& \ d_j = d_i\}$, i.e. Q is the set of all indices such that d_i is the first time this value appears in d. Note that Q_d is of size at most $|D|$. Let $Q^C_d = [n] \setminus Q_d$, let $R_d \subset D : \{d_i|i \in Q_d\}$ i.e. all the elements of D that appear at least once in the sequence d and let $R^C_d = D \setminus R_d$.

Let $y' = (y_i)_{i \in Q_d^C}$, $x_{Q_d} = (x_i)_{i \in Q_d}$, $\mathsf{Y}(R_d) = (\mathsf{Y}_d)_{d \in R_d}$ and $\mathsf{Y}(S_d) = (\mathsf{Y}_d)_{d \in S_d}$.

$$\mathbf{K}[\mathsf{X}_{[n]} | \mathsf{D}_{[n]} \mathsf{Z}]^x_{dz} = \sum_{y \in Y^{|D|}} \prod_{i \in Q_d} \delta[y_{d_i}]^{x_i} \mathbf{K}[\mathsf{Y}(D) | \mathsf{D}_{[n]} \mathsf{Z}]^y_{dz} \tag{161}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}(R_d) \mathsf{Y}(R_d^C) | \mathsf{D}_{Q_d} \mathsf{D}_{Q_d^C} \mathsf{Z}]^{x_{Q_d} y'}_{d_{Q_d} d^C_{Q_d} z} \tag{162}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}_{R_d} \mathsf{Y}_{R_d^C} | \mathsf{D}_{Q_d} \mathsf{D}_{Q_d^C} \mathsf{Z}]^{x_{Q_d} y'}_{dz} \tag{163}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}_{[n]} | \mathsf{D}_{[n]} \mathsf{Z}]^{x_{Q_d} y'}_{dz} \qquad \text{(using exchangeability)} \tag{164}$$

Note that

Only if: We aim to show that the sequences $\mathsf{Y}_{[n]}$ and $\mathsf{D}_{[n]}$ in a parallel potential outcomes submodel are exchangeable and deterministically reproducible. □

8 Appendix: Connection is associative

This will be proven with string diagrams, and consequently generalises to the operation defined by Equation **??** in other Markov kernel categories.

Define

$$\mathsf{I}_{K\cdot\cdot} := \mathsf{I}_K \setminus \mathsf{I}_L \setminus \mathsf{I}_J \tag{165}$$

$$\mathsf{I}_{KL\cdot} := \mathsf{I}_K \cap \mathsf{I}_L \setminus \mathsf{I}_J \tag{166}$$

$$\mathsf{I}_{K\cdot J} := \mathsf{I}_K \cap \mathsf{I}_J \setminus \mathsf{I}_L \tag{167}$$

$$\mathsf{I}_{KLJ} := \mathsf{I}_K \cap \mathsf{I}_L \cap \mathsf{I}_J \tag{168}$$

$$\mathsf{I}_{\cdot L\cdot} := \mathsf{I}_L \setminus \mathsf{I}_K \setminus \mathsf{I}_J \tag{169}$$

$$\mathsf{I}_{\cdot LJ} := \mathsf{I}_L \cap \mathsf{I}_J \setminus \mathsf{I}_K \tag{170}$$

$$\mathsf{I}_{\cdot\cdot J} := \mathsf{I}_J \setminus \mathsf{I}_K \setminus \mathsf{I}_L \tag{171}$$

$$\mathsf{O}_{K\cdot\cdot} := \mathsf{O}_K \setminus \mathsf{I}_N \setminus \mathsf{I}_J \tag{172}$$

$$\mathsf{O}_{KL\cdot} := \mathsf{O}_K \cap \mathsf{I}_L \setminus \mathsf{I}_J \tag{173}$$

$$\mathsf{O}_{K\cdot J} := \mathsf{O}_K \cap \mathsf{I}_J \setminus \mathsf{I}_L \tag{174}$$

$$\mathsf{O}_{KLJ} := \mathsf{O}_K \cap \mathsf{I}_L \cap \mathsf{I}_J \tag{175}$$

$$\mathsf{O}_{L\cdot} := \mathsf{O}_L \setminus \mathsf{I}_J \tag{176}$$

$$\mathsf{O}_{LJ} := \mathsf{O}_L \cap \mathsf{I}_J \tag{177}$$

Also define

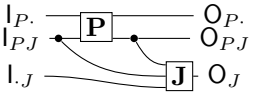
$$(\mathbf{P}, \mathsf{I}_P, \mathsf{O}_P) := \mathbf{K} \rightrightarrows \mathbf{L} \tag{178}$$

$$(\mathbf{Q}, \mathsf{I}_Q, \mathsf{O}_Q) := \mathbf{L} \rightrightarrows \mathbf{J} \tag{179}$$

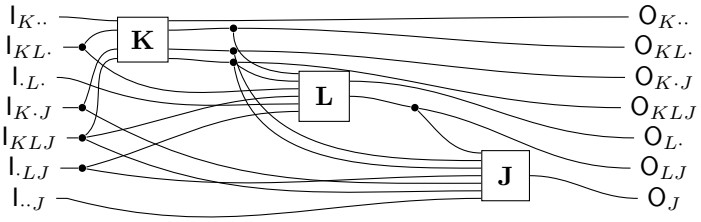
Then

$$(\mathbf{K} \rightrightarrows \mathbf{L}) \rightrightarrows \mathbf{J} = \mathbf{P} \rightrightarrows \mathbf{J} \tag{180}$$

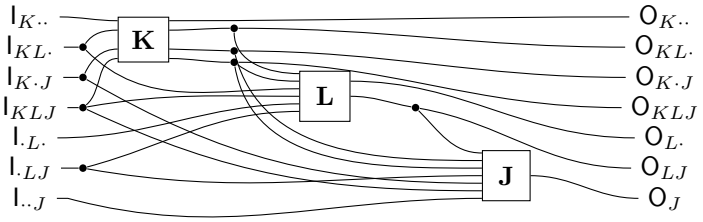
$$= \tag{181}$$



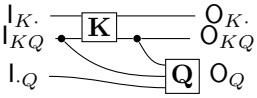
$$= \tag{182}$$



$$\overset{perm}{=} \tag{183}$$



$$= \tag{184}$$



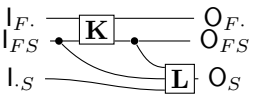
$$= \mathbf{K} \rightrightarrows (\mathbf{L} \rightrightarrows \mathbf{J}) \tag{185}$$

9 Appendix: String Diagram Examples

Recall the definition of *connection*:

Definition 9.1 (Connection)**.**

$$\mathbf{K} \rightrightarrows \mathbf{L} := \quad (186)$$

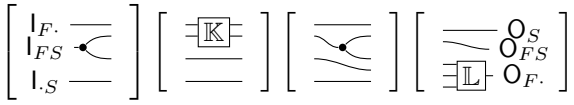


$$:= \mathbf{J} \quad (187)$$

$$\mathbf{J}^{zxw}_{yqr} = \mathbf{K}^{zx}_{yq}\mathbf{L}^{w}_{xqr} \quad (188)$$

Equation 186 can be broken down to the product of four Markov kernels, each of which is itself a tensor product of a number of other Markov kernels:

$$(\mathbf{J}, (\mathsf{I}_{F\cdot}, \mathsf{I}_{FS}, \mathsf{I}_{\cdot S}), (\mathsf{O}_{F\cdot}, \mathsf{O}_{FS}, \mathsf{O}_S)) = \quad (189)$$



(190)

10 Markov variable maps and variables form a Markov category

In the following, given *arbitrary measurable sets* $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$, a Markov kernel is a function $\mathbf{K} : X \times \mathcal{Y} \to [0, 1]$ such that

- For every $A \in \mathcal{Y}$, the function $x \mapsto \mathbf{K}(x, A)$ is $\mathcal{X}$-measurable
- For every $x \in X$, the function $A \mapsto \mathbf{K}(x, A)$ is a probability measure on $(Y, \mathcal{Y})$

Note that this is a more general definition than the one used in the main paper; the version in the main paper is the restriction of this definition to finite sets.

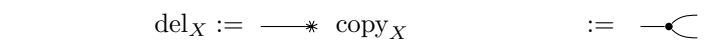
The *delta function* $\delta : X \to \Delta(\mathcal{X})$ is the Markov kernel defined by

$$\delta(x, A) = \begin{cases} 1 & x \in A \\ 0 & \text{otherwise} \end{cases} \quad (191)$$

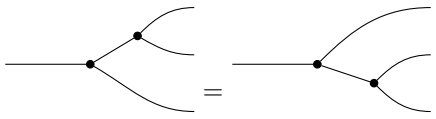
Fritz (2020) defines Markov categories in the following way:

Definition 10.1. A Markov category C is a symmetric monoidal category in which every object $X \in C$ is equipped with a commutative comonoid structure given by a comultiplication $\mathrm{copy}_X : X \to X \otimes X$ and a counit $\mathrm{del}_X : X \to I$, depicted in string diagrams as

$$\mathrm{del}_X := \quad \mathrm{copy}_X \quad := \quad (192)$$

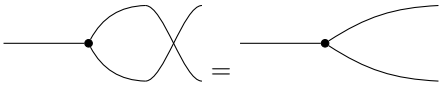


and satisfying the commutative comonoid equations



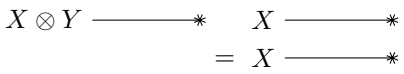
(193)

(194)

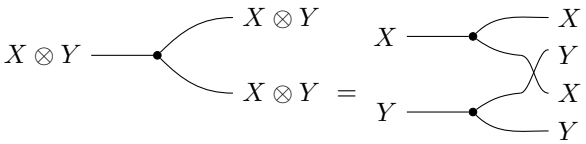


(195)

as well as compatibility with the monoidal structure

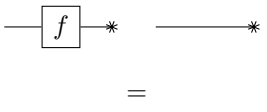


(196)



(197)

and the naturality of *del*, which means that



(198)

for every morphism f.

The category of labeled Markov kernels is the category consisting of labeled measurable sets as objects and labeled Markov kernels as morphisms. Given $\mathbf{K} : \mathsf{X} \to \Delta(\mathsf{Y})$ and $\mathbf{L} : \mathsf{Y} \to \Delta(\mathsf{Z})$, sequential composition is given by

$$\mathbf{KL} : \mathsf{X} \to \Delta(\mathsf{Z}) \tag{199}$$

$$\text{defined by } (\mathbf{KL})(x, A) = \int_Y \mathbf{L}(y, A)\mathbf{K}(x, dy) \tag{200}$$

For $\mathbf{K} : \mathsf{X} \to \Delta(\mathsf{Y})$ and $\mathbf{L} : \mathsf{W} \to \Delta(\mathsf{Z})$, parallel composition is given by

$$\mathbf{K} \otimes \mathbf{L} : (\mathsf{X}, \mathsf{W}) \to \Delta(\mathsf{Y}, \mathsf{Z}) \tag{201}$$

$$\text{defined by } \mathbf{K} \otimes \mathbf{L}(x, w, A \times B) = \mathbf{K}(x, A)\mathbf{L}(w, B) \tag{202}$$

The identity map is

$$\mathrm{Id}_{\mathsf{X}} : \mathsf{X} \to \Delta(\mathsf{X}) \tag{203}$$
$$\text{defined by} (\mathrm{Id}_X)(x, A) = \delta(x, A) \tag{204}$$

We take an arbitrary single element labeled set $I = (*, \{*\})$ to be the unit, which we note satisfies $I \otimes X = X \otimes I = X$ by Lemma **??**.

The swap map is given by

$$\mathrm{swap}_{\mathsf{X},\mathsf{Y}} : (\mathsf{X}, \mathsf{Y}) \to \Delta(\mathsf{Y}, \mathsf{X}) \tag{205}$$
$$\text{defined by} (\mathrm{swap}_{\mathsf{X},\mathsf{Y}})(x, y, A \times B) = \delta(x, B)\delta(y, A) \tag{206}$$

And we use the standard associativity isomorphisms for Cartesian products such that $(A \times B) \times C \cong A \times (B \times C)$, which in turn implies $(\mathsf{X}, (\mathsf{Y}, \mathsf{Z})) \cong ((\mathsf{X}, \mathsf{Y}), \mathsf{Z})$.

The copy map is given by

$$\mathrm{copy}_{\mathsf{X}} : \mathsf{X} \to \Delta(\mathsf{X}, \mathsf{X}) \tag{207}$$
$$\text{defined by} (\mathrm{copy}_X)(x, A \times B) = \delta_x(A)\delta_x(B) \tag{208}$$

and the erase map by

$$\mathrm{del}_{\mathsf{X}} : \mathsf{X} \to \Delta(*) \tag{209}$$
$$\text{defined by} (\mathrm{del}_X)(x, A) = \delta(*, A) \tag{210}$$
$$\tag{211}$$

Note that the category formed by taking the underlying unlabeled sets and the underlying unlabeled morphisms is identical to the category of measurable sets and Markov kernels described in Fong (2013); Cho and Jacobs (2019); Fritz (2020).

Theorem 10.2 (The category of labeled Markov kernels and labeled measurable sets is a Markov category)**.** *The category described above is a Markov category.*

Proof.

I'm not sure how to formally argue that it is monoidal and symmetric as the relevant texts I've checked all gloss over the functors with respect to which the relevant isomorphisms should be natural, but labels with products were intentionally made to act just like sets with cartesian products which are symmetric monoidal

Equations 193 to 198 are known to be satisfied for the underlying unlabeled Markov kernels. We need to show is that they hold given our stricter criterion of labeled Markov kernel equality; that the underlying kernels *and the label sets* match. It is sufficient to check the label sets only.

□

References

G. Chiribella, Giacomo D'Ariano, and P. Perinotti. Quantum Circuit Architecture. *Physical review letters*, 101:060401, September 2008. doi: 10.1103/PhysRevLett.101.060401.

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.

Panayiota Constantinou and A. Philip Dawid. EXTENDED CONDITIONAL INDEPENDENCE AND APPLICATIONS IN CAUSAL INFERENCE. *The Annals of Statistics*, 45(6):2618–2653, 2017. ISSN 0090-5364. URL `http://www.jstor.org/stable/26362953`. Publisher: Institute of Mathematical Statistics.

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.

R.P. Feynman. *The Feynman lectures on physics*. Le cours de physique de Feynman. Intereditions, France, 1979. ISBN 978-2-7296-0030-3. INIS Reference Number: 13648743.

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.

Tobias Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, August 2020. ISSN 0001-8708. doi: 10.1016/j.aim.2020.107239. URL `https://www.sciencedirect.com/science/article/pii/S0001870820302656`.

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`.

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`.

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.

Alfred Korzybski. *Science and sanity; an introduction to Non-Aristotelian systems and general semantics*. Lancaster, Pa., New York City, The International Non-Aristotelian Library Publishing Company, The Science Press Printing Company, distributors, 1933. URL http://archive.org/details/sciencesanityint00korz.

Finnian Lattimore and David Rohde. Causal inference with Bayes rule. *arXiv:1910.01510 [cs, stat]*, October 2019a. URL http://arxiv.org/abs/1910.01510. arXiv: 1910.01510.

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

David Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59(1): 5–30, March 1981. ISSN 0004-8402. doi: 10.1080/00048408112340011. URL https://doi.org/10.1080/00048408112340011.

Karl Menger. Random Variables from the Point of View of a General Theory of Variables. In Karl Menger, Bert Schweizer, Abe Sklar, Karl Sigmund, Peter Gruber, Edmund Hlawka, Ludwig Reich, and Leopold Schmetterer, editors, *Selecta Mathematica: Volume 2*, pages 367–381. Springer, Vienna, 2003. ISBN 978-3-7091-6045-9. doi: 10.1007/978-3-7091-6045-9_31. URL https://doi.org/10.1007/978-3-7091-6045-9_31.

Scott Mueller, Ang Li, and Judea Pearl. Causes of Effects: Learning individual responses from population data. *arXiv:2104.13730 [cs, stat]*, May 2021. URL http://arxiv.org/abs/2104.13730. arXiv: 2104.13730.

Paul F. Christiano. EDT vs CDT, September 2018. URL https://sideways-view.com/2018/09/19/edt-vs-cdt/.

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

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

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.

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.

Eyal Shahar. The association of body mass index with health outcomes: causal, inconsistent, or confounded? *American Journal of Epidemiology*, 170(8):957–958, October 2009. ISSN 1476-6256. doi: 10.1093/aje/kwp292.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, October 2000. ISSN 1573-7470. doi: 10.1023/A:1018912507879. URL https://doi.org/10.1023/A:1018912507879.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, July 2002.

J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Theory of games and economic behavior. Princeton University Press, Princeton, NJ, US, 1944.

Abraham Wald. *Statistical decision functions*. Statistical decision functions. Wiley, Oxford, England, 1950.

Paul Weirich. Causal Decision Theory. 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/decision-causal/.

Appendix: