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

David Johnston

October 27, 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 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

[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.

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 Why are variables functions?

Our theory is mainly concerned with *Markov kernels*, which are functions that map from measurable sets to probability distributions. On their own, Markov kernels are just abstract mathematical objects, but we want to use them to make *models* which aim to explain some part of the world. To make a Markov kernel into a model, we annotate it with a set of variables. The role of variables is to point to some parts of the real world. The model relates variables to one another "in the world of maths", and the variables enable this to be interpreted as a claim about how some parts of the real world relate to one another.

We will consider an example with no probabilistic aspects. Newton's second law, $F = ma$, relates "variables" F, m and a. As Feynman (1979) noted, this law cannot serve as a definition of force, nor mass or acceleration. It relates three things – force, mass and acceleration – that we are presumed to know something about already. In particular, this law suggests that there is some way to take ane object and determine the total forces acting on it, and some other way to determine its mass and acceleration, and the force should agree with the product of the mass and acceleration.

As it happens, physicists know a great deal about determining the forces on an object, its mass and its acceleration, and furthermore know a fair bit about what an "object" is. Nevertheless, should we ever want to check the mass on an actual object we have to leave the world of mathematics in which our model resides, consult the actual object about its mass, and return with a number inside our mathematical world.

These kinds of variables are odd. They have a well-defined mathematical codomain, but their domain is a "part of the real world". An important feature of them is that if we compose them with a function, we obtain a new variable. If "m" takes an object in the world and maps it to a mass, then "$2m$" is a variable that, when given the same object as m, returns twice the value that m does.

We cannot, however, compose functions with variables – I cannot multiply an object in the real world by 2, as there is no Cartesian product between real numbers and things in the world on which to define multiplication.

We can't operate mathematically with such things, but we want to work with a model of them that does not contradict their actual behaviour. We can construct such a model

Variables are what we use to explain what our models are actually modelling.

Our intention is to clarify various "map-territory" issues related to causal inference. We start with a discussion of variables. Variables are often where we specify which real things our models are supposed to correspond to, and deserve special attention because causal inference problems often feature "unobserved" variables that play important roles.

A standard formal definition of random variables is that they are functions from some measurable sample space $(\Omega, \mathcal{F})$ equipped with a probability measure $\mathbb{P}$ to some codomain $(X, \mathcal{X})$. This is also frequently relaxed to drop the requirement of a probability measure $\mathbb{P}$. This is how we will define variables, and will refer to them strictly as *formal variables.* However, we will also hold to the typical practice of avoiding explicit definition of a sample space, and we will say a few words about why this model is used.

Typically, variables are often defined with reference to:

- The things that they are supposed to represent
- Their codomain

For example, Pearl (2009) offers the same definition of formal variables as we do, but also explains:

> 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.

We will call variables defined with reference to specific measurements *observable variables.* How should we model variables of this kind? Suppose we have an observable variable X – i.e. X represents an act of measurement ("the measurement of my height in centimetres I am about to perform") and has a specified codomain (a natural number). Suppose we can consult X on multiple occasions – say, first to write it in this paper (185cm) and secondly to discuss with my daughter later today. I cannot consult X before I measure my height, but once I have done so, in order for X to be well-defined, it must surely have the same value every time it is consulted.

Menger (2003), under the name of *qualitative statistical random variable*, offers the following definition of a variable that satisfies this condition. A qualitative statistical random variable is:

- A set of ordered pairs (x, y), where xs are attributes, acts of measurement or inquiries and ys may take values in a specified codomain Y

- Each x corresponds to exactly one y

Menger offers the examples of a variable that takes for x people in Chicago and returns as y their height in metres. The condition each x corresponds to exactly one y means that, if we go and do the experiment and measure the height of everybody in Chicago, each person will be assigned exactly one canonical height, and every time we query the height of a person x we get the same height y.

Our first condition for a model of observable variables is, if the value of an observable variable appears in multiple places, it must in all cases take the same value.

Secondly, it is often useful to consider "functions of variables". For example, my daughter might want to know if she is more than half of my height. If my height is X, then we will want to know what half my height is. We can consider this a second variable H, whose codomain is multiples of $\frac{1}{2}$, with the property that whatever value y is the "actual" value of X, the value of H must be $\frac{y}{2}$. This induces a binary relation between the codomains of X and H consisting of the valid pairs of values $(y, \frac{y}{2})$.

For theoretical tractability, however, we require formal definitions of observable variables.

$$\mathsf{X} = \mathsf{X} \tag{1}$$

An important question for our purposes is: when we define vague variables, what relationships should we understand them to have? This question can be illustrated with an example from the causal modelling framework of *structural causal models*. This framework assigns a function to each variable. For example

$$\mathsf{X} = f_0(\xi_{\mathsf{X}}) \tag{2}$$
$$\mathsf{Y} = f_1(\mathsf{X}, \xi_{\mathsf{Y}}) \tag{3}$$

We may then *intervene* on some or all variables. An intervention on X, in this case, is an operation that generates a new set of structural equations identical to the previous set, except with the intervened variable having its equation modified in some manner (in this example, this could involve replacing Equation 2 with $\mathsf{X} = 1$).

Are the function assignments in a set of structural equations models of measurement processes? That is, if X is defined as a vague variable, is the function $\mathsf{X} \circ \xi_{\mathsf{X}}$ the model of the measurment process of interest? We think that the answer is *no*. Interventions are typically meant to model actions that cause X to "actually take on a different value", not actions that alter the way we turn measurements into values. We could also say: if we intervene on X, the way we determine the value of X from a measurement stays the same, but the result we get might be different. The *definition of the variable* X is a higher-order commitment, which must be respected by any model of X, while the definition of the *structural equation associated with* X is a lower-order commitment which may be violated if we consider acting to change X.

Here we investigate an approach to probability modelling in which variables are higher-order commitments. That is, any valid model must satisfy constraints imposed by the definitions of variables. We show that this approach reduces to the standard approach under appropriate conditions, and is also a practical approach for dealing with the multiplicity of probability models that we often find ourselves considering when doing causal inference.

2.2 Probability, variables and composition

Throughout this paper, we will assume all measurable sets are finite sets. This is because it makes explanations simpler and because it is easy to show that conditional probabilities exist in this setting (Lemma 2.18).

We assume that there is some measurable sample space Ω and that all variables are measurable functions defined on Ω. It is also often standard to assume that we have a *probability space* $(\mathbb{P}, (\Omega, \mathcal{F}))$, where $\mathbb{P}$ is a σ-additive measure on $(\Omega, \mathcal{F})$ with $\mathbb{P}(\Omega) = 1$. Given such a probability space, the normal way to define the probability distribution of a particular random variable is via the *pushforward measure*. Given $(\mathbb{P}, (\Omega, \mathcal{F}))$ and $f_{\mathsf{X}} : (\Omega, \mathcal{F}) \to (X, \mathcal{X})$, the pushforward measure is $\mathbb{P}^{\mathsf{X}}(A) := \mathbb{P}(\mathsf{X}^{-1}(A))$ for $A \in \mathcal{X}$.

We use a different approach to defining "the distribution of X". Rather than defining it directly via the pushforward measure, we hold that, firstly, any contradictions by our definitions of random variables must be given probability 0, a property called *consistency*, and secondly $\mathbb{P}^{\mathsf{X}}$ can be obtained from $\mathbb{P}^{\mathsf{XY}}$ by marginalising over Y. Together with a probability $\mathbb{P}$ defined on the entire sample space, these recover the pushforward rule.

We will use the example of truncated factorisation to explain the motivation behind our approach. Consider "truncated factorisation". Suppose we have a causal Bayesian network $(\mathbb{P}^{\mathsf{XYZ}}, \mathcal{G})$ where $\mathcal{G}$ is a Directed Acyclic Graph that contains the edges $\mathsf{X} \longrightarrow \mathsf{Y}$ and $\mathsf{X} \longleftarrow \mathsf{Z} \longrightarrow \mathsf{Y}$. Then the result of "setting X to x" is represented by a new probability measure $\mathbb{P}_x$ that is required to obey truncated factorisation (Pearl, 2009, page 24):

$$\mathbb{P}_x^{\mathsf{XYZ}}(x', y, z) = \mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}(y|x, z)\mathbb{P}^{\mathsf{Z}}(z)[\![x = x']\!] \qquad (4)$$

Equation 4 embodies three assumptions. First, when we set X to x, then $\mathsf{X} \overset{a.s.}{=} x$. Second, when we set X to x, then Z carries on as before. Finally, Y given X and Z also carries on as before. These assumptions are not all equal in stature: the condition on the distribution of X is absolutely crucial: this is what it means to set X to x. The other two might be good assumptions if this causal Bayesian network happens to be good for our purposes.

However, there might be other assumptions that are more forceful than these latter two. For example, if X and Z happened to actually be the same random variables – the same thing in the world that we go and look at – then we absolutely must have $\mathsf{X} \overset{a.s.}{=} \mathsf{Z}$. The standard method for determining $\mathbb{P}_x^{\mathsf{XYZ}}$ will normally ensure that this condition is satisfied; that is, taking some $\mathbb{P}_x$ and compute the pushforward under $(\mathsf{X}, \mathsf{Y}, \mathsf{Z})$ will ensure $\mathsf{X} \overset{a.s.}{=} \mathsf{Z}$ if indeed $\mathsf{X} = \mathsf{Z}$.

However, $\mathsf{X} \overset{a.s.}{=} \mathsf{Z}$ cannot in general be satisfied at the same time as Equation 4 for all x. Indeed, if x may take more than one value, these two conditions cannot be simultaneously satisfied for at least one value of x.

So we have one critical assumption – that $\mathsf{X} \overset{a.s.}{=} x$ – from Equation 4, and another critical assumption – that $\mathsf{X} \overset{a.s.}{=} \mathsf{Z}$ – from the standard definition of what "$\mathbb{P}_x^{\mathsf{XYZ}}$" means, *and* we know that Equation 4 cannot actually be satisfied. This is obviously a mixed up situation. The condition of *consistency*, which we introduce, addresses this problem. In this case, consistency demands $\mathsf{X} \overset{a.s.}{=} x$ and $\mathsf{X} \overset{a.s.}{=} \mathsf{Z}$ together, as we will show. Assumptions like Equation 4 can be added as "lower order" demands of our probability model, but if they violate consistency then we must abandon them.

The assumption that $\mathsf{Z} = \mathsf{X}$ might seem forced, however we can consider a very similar situation if $\mathsf{Z} = (\mathsf{H}, \mathsf{W})$, representing the height in metres and weight in kilograms of a particular person, and X represents their body mass index. In this case the causal structure we proposed is not original to us – it appears in Shahar (2009). However, it is the case $\mathsf{X} = \frac{\mathsf{W}}{\mathsf{H}^2}$ and this also imposes a constraint that cannot be satisfied at the same time as 4.

The condition of consistency allows us to check when non-standard products like Equation 4 yield "well-formed" probability models on the listed variables. We offer some sufficient conditions for probability models to be well-formed, which includes the case where the variables in question are surjective and *variationally independent*, and if we have a strictly positive model over the same variables we already know to be well-formed.

We also show that the standard approach of defining a sample space model and defining marginals and conditionals via push-forwards is safe, in the sense that if the sample space model is well-formed then the marginals are well-formed and conditionals can always be chosen to be well-formed.

2.3 Probability and composition without variables: Markov categories

Markov categories are abstract categories that represent models of the flow of information. Operations like Equation 4 are expressible as abstract compositions in Markov categories, and may be represented with string diagrams developed for reasoning about objects in the category. Valid proofs using string diagrams correspond to valid theorems in *any* Markov category, though we will limit our attention to the category of finite sets and Markov kernels in this paper. The main drawback of Markov categories is that, as they exist at the moment, they have no notion of "variables". More comprehensive introductions to Markov categories can be found in Fritz (2020); Cho and Jacobs (2019).

Rather than explain Markov categories in the abstract, we will introduce string diagrams with reference to how they represent stochastic maps and finite sets (though see Appendix 10). Given measurable sets $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$, a Markov kernel or stochastic map is a map $\mathbf{K} : X \times \mathcal{Y} \to [0, 1]$ such that

- The map $x \mapsto \mathbf{K}(x, A)$ is $\mathcal{X}$-measurable for every $A \in \mathcal{Y}$

- The map $A \mapsto \mathbf{K}(x, A)$ is a probability measure for every $x \in X$

Where X and Y are finite sets with the discrete σ-algebra, we can represent a Markov kernel $\mathbf{K}$ as a $|X| \times |Y|$ matrix where $\sum_{y \in Y} \mathbf{K}_x^y = 1$ for every $x \in X$. We will give Markov kernels the signature $\mathbf{K} : X \rightarrowtriangle Y$ to indicate that they map from X to probability distributions on Y.

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

$$\mathbf{K} := \;\; -\boxed{\mathbf{K}}- \tag{5}$$

$$\mathbf{P} := \lhd\!\mathbf{P}- \tag{6}$$

Two Markov kernels $\mathbf{L} : X \rightarrowtriangle Y$ and $\mathbf{M} : Y \rightarrowtriangle Z$ have a product $\mathbf{LM} : X \rightarrowtriangle Z$ given by the matrix product $\mathbf{LM}_x^z = \sum_y \mathbf{L}_x^y \mathbf{M}_y^z$. Graphically, we write represent by joining wires together:

$$\mathbf{LM} := \;\; -\boxed{\mathbf{K}}-\boxed{\mathbf{M}}- \tag{7}$$

The Cartesian product $X \times Y := \{(x, y)|x \in X, y \in Y\}$. Given kernels $\mathbf{K} : W \rightarrowtriangle Y$ and $\mathbf{L} : X \rightarrowtriangle Z$, the tensor product $\mathbf{K} \otimes \mathbf{L} : W \times X \rightarrowtriangle Y \times Z$ is defined by $(\mathbf{K} \otimes \mathbf{L})_{(w,x)}^{(y,z)} := K_w^y L_x^z$ and represents applying the kernels in parallel to their inputs.

The tensor product is represeted by drawing kernels in parallel:

$$\mathbf{K} \otimes \mathbf{L} := \begin{array}{c} W\,\boxed{\mathbf{K}}\,Y \\ X\,\boxed{\mathbf{L}}\,Z \end{array} \tag{8}$$

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)). A diagram describes products and tensor products of Markov kernels, which are expressed according to the conventions described above. 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.

A description of these kernels follows.

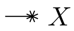
The identity map $\mathrm{id}_X : X \rightarrowtriangle X$ defined by $(\mathrm{id}_X)_x^{x'} = [\![x = x']\!]$, where the iverson bracket $[\![\cdot]\!]$ evaluates to 1 if $\cdot$ is true and 0 otherwise, is a bare line:

$$\mathrm{id}_X := \;\; X - X \tag{9}$$

We choose a particular 1-element set $\{*\}$ that acts as the identity in the sense that $\{*\} \times A = A \times \{*\} = A$ for any set A. The erase map $\mathrm{del}_X : X \rightarrowtriangle \{*\}$

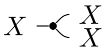
defined by $(\mathrm{del}_X)^*_x = 1$ is a Markov kernel that "discards the input" (we will later use it for marginalising joint distributions). It is drawn as a fuse:

$$\mathrm{del}_X := \qquad (10)$$



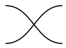
The copy map $\mathrm{copy}_X : X \rightarrowtriangle X \times X$ defined by $(\mathrm{copy}_X)^{x',x''}_x = [\![x = x']\!][\![x = x'']\!]$ is a Markov kernel that makes two identical copies of the input. It is drawn as a fork:

$$\mathrm{copy}_X := \qquad (11)$$



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

$$\mathrm{swap}_X := \qquad (12)$$

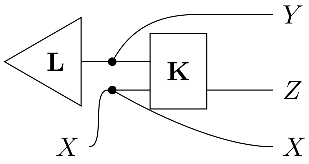


Because we anticipate that the graphical notation will be unfamiliar to many, we will also include translations to more familiar notation.

2.4 Truncated factorisation with Markov kernels

The Markov kernels introduced in the previous section can be though of as "conditional probability distributions without variables". We can use these to represent an operation very similar to Equation 4. Note that $P^{\mathsf{Y}|\mathsf{XZ}}$ must be represented by a Markov kernel $\mathbf{K} : X \times Z \rightarrowtriangle Y$ and $\mathbb{P}^{\mathsf{Z}}$ by a Markov kernel $\mathbf{L} \in \Delta(Z)$. Then we can define a Markov kernel $\mathbf{M} : X \rightarrowtriangle X \times Z$ representing $x \mapsto \mathbb{P}^{\mathsf{YZ}}_x(y, z)$ by

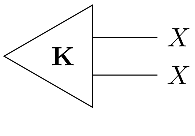
$$\mathbf{M} := \qquad (13)$$



There is, however, a key difference between Equation 13 and Equation 4: the Markov kernels in the latter equation describe the distribution of particular variables, while the former equation describes Markov kernels only.

To illustrate why we need variables, consider an arbitrary Markov kernel $\mathbf{K} : \{*\} \rightarrowtriangle \Delta(X \times X)$. We could draw this:

$$\mathbf{K} := \qquad (14)$$



We label both wires with the set X. However, say $X = \{0, 1\}$. Then $\mathbf{K}$ could be the kernel $\mathbf{K}^{x_1,x_2} = \llbracket x_1 = 0 \rrbracket \llbracket x_2 = 1 \rrbracket$. In this case, both of its outputs must represent *different* variables, despite taking values in the same set. On the other hand, if $\mathbf{K}^{x_1,x_2} = 0.5 \llbracket x_1 = x_2 \rrbracket$ then both outputs coudl represent the same variable, because they are deterministically the same, or they could represent different variables that happen to be equal. We need some way to distinguish the two cases.

2.5 Composition and probability with variables

Our goal is to define a category of "finite sets and Markov kernels with variables". Introducing variables requires an assumption of consistency, which we don't know how to express in category theoretic terms. Our approach is to define a category of Markov kernels with variables that may or may not be consistent, which we will need to check for the resulting models. Because the consistency assumption is not expressed category theoretically, many proofs in this section only apply to our chosen setting of finite sets.

Definition 2.1 (Variable)**.** Given a *sample space* Ω, a variable f_{X} is a function $\Omega \to A$ where A is a vector space. We will also refer to the associated Markov kernel $\mathsf{X} : \Omega \rightarrowtriangle A$ as a variable, where $\mathsf{X}_x^a = \llbracket a = f_{\mathsf{X}}(x) \rrbracket$.

We define the *product* of two variables as follows:

- **Product:** Given variables $\mathsf{W} : \Omega \rightarrowtriangle A$ and $\mathsf{V} : \Omega \rightarrowtriangle B$, the product is defined as $(\mathsf{W}, \mathsf{V}) = \text{copy}_{\Omega}(\mathsf{W} \otimes \mathsf{V})$

The *unit* variable is the erase map $\mathsf{I} := \text{del}_{\Omega}$, with $(\mathsf{I}, \mathsf{X}) = (\mathsf{X}, \mathsf{I}) = \mathsf{X}$ (up to isomorphism) for any X.

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

Definition 2.2 (Indexed Markov kernel)**.** Given variables $\mathsf{X} : \Omega \to A$ and $\mathsf{Y} : \Omega \to B$, an indexed Markov kernel $\mathbf{K} : \mathsf{X} \rightarrowtriangle \mathsf{Y}$ is a triple $(\mathbf{K}', \mathsf{X}, \mathsf{Y})$ where $\mathbf{K}' : A \rightarrowtriangle B$ 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:

$$\mathbf{K} := \begin{matrix} \mathsf{A}_1 \\ \mathsf{A}_2 \end{matrix} - \boxed{\mathbf{K}} - \begin{matrix} \mathsf{B}_1 \\ \mathsf{B}_2 \end{matrix} \tag{15}$$

or

$$\mathbf{K} = (\mathsf{A}_1, \mathsf{A}_2) - \boxed{\mathbf{K}[\mathbb{L}]} - (\mathsf{B}_1, \mathsf{B}_2) \tag{16}$$

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

Similarly, the tensor product of $\mathbf{K} : \mathsf{X} \rightarrowtriangle \mathsf{Y}$ and $\mathbf{L} : \mathsf{W} \rightarrowtriangle \mathsf{Z}$ 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 del_X, copy_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 $\mathbf{K} : \mathsf{X} \rightarrowtriangle \mathsf{X}$ with $\mathbf{K}_a'^b = 0.5$ 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 2.3 (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:

$$f_\mathsf{X}^{-1}(a) \cap f_\mathsf{Y}^{-1}(b) = \emptyset \implies \left(\mathbf{K}_a'^b = 0\right) \vee \left(f_\mathsf{X}^{-1}(a) = \emptyset\right) \tag{17}$$

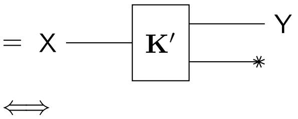
A *probability model* is a model where the underlying kernel $\mathbf{K}'$ has the unit I as the domain. We use the font $\mathbb{K}$ to distinguish models from arbitrary indexed Markov kernels.

Consistency implies that for any $\mathbb{K} : \mathsf{X} \rightarrowtriangle \mathsf{Y}$, if $f_\mathsf{Y} = g \circ f_\mathsf{X}$ then $\mathbb{K}_x^{g(x)} = 1$. A particularly simple case of this is a model $\mathbb{L} : \mathsf{X} \rightarrowtriangle \mathsf{X}$, 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 2.4 (Marginal distribution)**.** Given a model $\mathbb{K} : \mathsf{X} \rightarrowtriangle (\mathsf{Y}, \mathsf{Z})$, the marginal distribution of Y, written $\mathbb{K}^{\mathsf{Y}|\mathsf{X}}$, is obtained by marginalising over Z:

$$\mathbb{K}^{\mathsf{Y}|\mathsf{X}} := \tag{18}$$



$$\iff \tag{19}$$

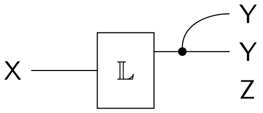
$$(\mathbb{K}^{\mathsf{Y}|\mathsf{X}})_x^y = \sum_{z \in Z} \mathbf{K}_x'^{yz} \tag{20}$$

Definition 2.5 (Disintegration)**.** Given a model $\mathbb{K} : \mathsf{X} \rightarrowtriangle (\mathsf{Y}, \mathsf{Z})$, a disintegration $\mathbb{L} : (\mathsf{X}, \mathsf{Y}) \rightarrowtriangle \mathsf{Z}\ \mathsf{Y}$, 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 2.6 (Output copies of the same variable are identical)**.** *For any* $\mathbf{K} : \mathsf{X} \rightarrowtriangle (\mathsf{Y}, \mathsf{Y}, \mathsf{Z})$*,* $\mathbf{K}$ *is a model iff there exists some* $\mathbb{L} : \mathsf{X} \rightarrowtriangle (\mathsf{Y}, \mathsf{Z})$ *such that*

$$\mathbf{K} = \tag{21}$$



$$\iff \tag{22}$$

$$\mathbf{K}'^{y,y',z}_x = \llbracket y = y' \rrbracket \mathbf{L}'^{y,z}_x \tag{23}$$

$$\tag{24}$$

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

$$(\mathsf{X}, \mathsf{Y}, \mathsf{Y}, \mathsf{Z})^{x,y,y',z}_\omega = \llbracket f_\mathsf{Y}(\omega) = y \rrbracket \llbracket f_\mathsf{Y}(\omega) = y' \rrbracket (\mathsf{X}, \mathsf{Z})^{x,z}_\omega \tag{25}$$

$$= \llbracket y = y' \rrbracket \llbracket f_\mathsf{Y}(\omega) = y \rrbracket (\mathsf{X}, \mathsf{Z})^{x,z}_\omega \tag{26}$$

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

$$\mathbf{K}'^{y,y',z}_x = \llbracket y = y' \rrbracket \mathbf{L}'^{y,z}_x \tag{27}$$

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

$$\llbracket y = y' \rrbracket \mathbb{L}'^{y,z}_x > 0 \implies y = y' \wedge \mathbb{L}'^{y,z}_x > 0 \tag{28}$$

$$\implies \left(f_\mathsf{X}^{-1}(x) = \emptyset\right) \vee \left(f_\mathsf{X}^{-1}(x) \cap f_\mathsf{Y}^{-1}(y) \cap f_\mathsf{Y}^{-1}(y) \cap f_\mathsf{Z}^{-1}(z) \neq \emptyset\right) \tag{29}$$

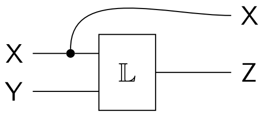
$$\tag{30}$$

□

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

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

$\mathbf{K} =$  (31)

$$\iff \tag{32}$$

$$\mathbf{K}'^{x',z}_{x,y} = [\![x = x']\!]\mathbf{L}^{z}_{\prime x,y} \tag{33}$$

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

$$\begin{align}
(\mathsf{X}, \mathsf{Y}, \mathsf{Y}, \mathsf{Z})^{x,y,y',z}_{\omega} &= [\![f_{\mathsf{Y}}(\omega) = y]\!][\![f_{\mathsf{Y}}(\omega) = y']\!](\mathsf{X}, \mathsf{Z})^{x,z}_{\omega} \tag{34}\\
&= [\![y = y']\!][\![f_{\mathsf{Y}}(\omega) = y]\!](\mathsf{X}, \mathsf{Z})^{x,z}_{\omega} \tag{35}
\end{align}$$

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

$$\mathbf{K}'^{x',z}_{x,y} = [\![x = x']\!]\mathbf{L}'^{z}_{x,y} \tag{36}$$

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

$$\begin{align}
[\![x = x']\!]\mathbb{L}'^{z}_{x,y} > 0 &\implies x = x' \wedge \mathbb{L}'^{z}_{x,y} > 0 \tag{37}\\
&\implies \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) = \emptyset\right) \vee \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) \cap f_{\mathsf{Z}}^{-1}(z) \neq \emptyset\right) \tag{38}\\
& \tag{39}
\end{align}$$

□

Consistency along with the notion of marginal distributions implies that, given some X and some $\mathbb{K} : \mathsf{Y} \rightarrowtriangle \mathrm{Id}_{\Omega}$, the pushforward $\mathbb{K}\mathbb{X}$ is the unique model $\mathsf{Y} \rightarrowtriangle \mathsf{X}$ that can be paired (Definition 2.9) with $\mathbb{K}$. This is shown in Lemma 2.10.

Lemma 2.8 (Uniqueness of models with the sample space as a domain)**.** *For any* $\mathsf{X} : \Omega \to A$*, there is a unique model* $\mathbb{X} : Id_{\Omega} \rightarrowtriangle \mathsf{X}$ *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 $\omega \in \Omega$, $a \in A$

$$\begin{align}
\max_{\omega \in \Omega} (\mathrm{Id}_{\Omega}, \mathsf{X})^{\omega',a}_{\omega} &= \max_{\omega \in \Omega} [\![\omega = \omega']\!][\![\omega = f_{\mathsf{X}}(a)]\!] \tag{40}\\
&= [\![\omega = f_{\mathsf{X}}(a)]\!] \tag{41}\\
&= \mathbf{X}^{a}_{\omega} \tag{42}
\end{align}$$

Thus $\mathbb{X}$ satisfies consistency.

Suppose there were some $\mathbb{K} : \mathrm{Id}_\Omega \rightarrowtriangle \mathsf{X}$ 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

$$\max_{\omega\in\Omega}(\mathrm{Id}_\Omega, \mathsf{X})^{\omega',a}_{\omega} = \max_{\omega\in\Omega}[\![\omega = \omega']\!][\![\omega = f_\mathsf{X}(b)]\!] \tag{43}$$

$$= [\![\omega = f_\mathsf{X}(b)]\!] \tag{44}$$

$$= 0 \tag{45}$$

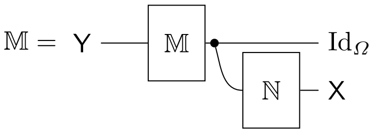
$$< \mathbb{K}^b_\omega \tag{46}$$

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

Definition 2.9 (Pairing)**.** Two models $\mathbb{K} : \mathsf{X} \rightarrowtriangle \mathsf{Y}$ and $\mathbb{L} : \mathsf{X} \rightarrowtriangle \mathsf{Z}$ can be *paired* if there is some $\mathbb{M} : \mathsf{X} \rightarrowtriangle (\mathsf{Y}, \mathsf{Z})$ such that $\mathbb{K} = \mathbb{M}^{\mathsf{Y}|\mathsf{X}}$ and $\mathbb{L} = \mathbb{M}^{\mathsf{Z}|\mathsf{X}}$.

Lemma 2.10 (Pushforward model)**.** *Given any model $\mathbb{K} : \mathsf{Y} \rightarrowtriangle \mathrm{Id}_\Omega$ and any X, there is a unique $\mathbb{L} : \mathsf{Y} \rightarrowtriangle \mathsf{X}$ that can be paired with $\mathbb{K}$, and it is given by $(\mathbf{L}^a_b = \sum_{\omega\in f^{-1}_{\mathsf{X}}(a)} \mathbf{K}^\omega_b$.*

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



(47)

By Corollary **??**, there is only one model $\mathbb{N} : \mathrm{Id}_\Omega \rightarrowtriangle \mathsf{X}$ 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 $a \in A$, $b \in B$

$$(\mathbb{K}\mathbb{X})^a_b = \sum_{\omega\in\Omega} \mathbb{P}^\omega_b \mathsf{X}^a_\omega \tag{48}$$

$$= \sum_{\omega\in\Omega} \mathbb{P}^\omega_b [\![a = f_\mathsf{X}(\omega)]\!] \tag{49}$$

$$= \sum_{\omega\in f^{-1}(a)} \mathbb{P}^\omega_b \tag{50}$$

□

Corollary 2.11 (Pushforward probability model)**.** *Given any probability model* $\mathbb{P} : \mathsf{I} \rightarrowtriangle Id_{\Omega}$*, there is a unique model* $\mathbb{P}^{\mathsf{X}} : \mathsf{I} \rightarrowtriangle \mathsf{X}$ *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 2.10 to a model $\mathbb{P} : \mathsf{I} \rightarrowtriangle \mathrm{Id}_{\Omega}$. □

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

$$f_{\mathsf{Y}}(\Omega) \times f_{\mathsf{X}}(\Omega) = \{(f_{\mathsf{Y}}(\omega), f_{\mathsf{X}}(\omega))|\omega \in \Omega\} \tag{51}$$

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

Lemma 2.13 (Consistency via variation conditional independence)**.** *Given an indexed Markov kernel* $\mathbf{K} : \mathsf{X} \rightarrowtriangle \mathsf{Y}$ *with* $\mathsf{X} : \Omega \rightarrowtriangle X$ *and* $\mathsf{Y} : \Omega \rightarrowtriangle Y$*, 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 2.6 and 2.7 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 2.14 (Consistency via positive models)**.** *Given a model* $\mathbb{K} : \mathsf{X} \rightarrowtriangle (\mathsf{Y}, \mathsf{Z})$*, if an indexed Markov kernel* $\mathbf{L} : (\mathsf{X}, \mathsf{Y}) \rightarrowtriangle \mathsf{Z}$ *has the property* $\mathbf{K}^{\prime yz}_{x} = 0 \implies \mathbf{L}^{\prime z}_{xy} = 0$ *then* $\mathbf{L}$ *is also a model.*

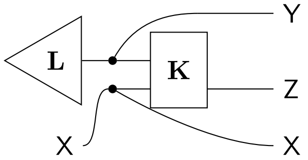
Proof. Because $\mathbb{K}$ is a model,

$$\begin{aligned}
\mathbf{L}^{\prime z}_{xy} > 0 &\implies \mathbf{K}^{\prime yz}_{x} > 0 && (52)\\
&\implies \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) \cap f_{\mathsf{Z}}^{-1}(z) \neq \emptyset\right) \vee \left(f_{\mathsf{X}}^{-1}(x) = \emptyset\right) && (53)\\
&\implies \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) \cap f_{\mathsf{Z}}^{-1}(z) \neq \emptyset\right) \vee \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) = \emptyset\right) && (54)
\end{aligned}$$

□

2.6 Truncated factorisation with variables

At this point, we can represent Equation 4 using models. Suppose $P^{\mathsf{Y}|\mathsf{X}\mathsf{Z}}$ is an model $\mathbb{K} : (\mathsf{X}, \mathsf{Z}) \rightarrowtriangle \mathsf{Y}$ and $\mathbb{P}^{\mathsf{Z}}$ an model $\mathbb{L} : \{*\} \rightarrowtriangle \mathsf{Z}$. Then we can define an indexed Markov kernel $\mathbf{M} : \mathsf{X} \rightarrowtriangle \mathsf{X}, \mathsf{Z}$ representing $x \mapsto \mathbb{P}_x^{\mathsf{YZ}}(y, z)$ by

$\mathbf{M} :=$  (55)

Equation 55 is almost identical to Equation 13, except it now specifies which variables each measure applies to, not just which sets they take values in. Like the original Equation 4, 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 2.6 to 2.14, although note that 2.13 and 2.14 are sufficient conditions, not necessary ones.

2.7 Sample space models and submodels

Instead of trying to assemble probability models as in Equation 55, 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 $\mathbf{K} : \mathsf{X} \rightarrowtriangle \mathrm{Id}_{\Omega}$. 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 2.15 (Existence of sample space model)**.** *Given any model $\mathbb{K} : \mathsf{X} \rightarrowtriangle \mathsf{Y}$, there is a sample space model $\mathbb{L} : \mathsf{X} \rightarrowtriangle Id_{\Omega}$ such that, defining $\mathbb{Y} := (\mathsf{Y}, Id_{\Omega}, \mathsf{Y})$, $\mathbb{L}\mathbb{Y} = \mathbb{K}$.*

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

$$\mathbf{L}_a'^{\omega} = \begin{cases} 0 & \text{if } f_{\mathsf{Y}}^{-1}(b) \cap f_{\mathsf{X}}^{-1}(a) = \emptyset \\ \mathbf{K}_a'^{b}[\![\omega = \omega_b]\!] & \text{for some } \omega_b \in f_{\mathsf{Y}}^{-1}(b) \text{ if } f_{\mathsf{X}}^{-1}(a) = \emptyset \\ \mathbf{K}_a'^{b}[\![\omega = \omega_{ab}]\!] & \text{for some } \omega_{ab} \in f_{\mathsf{Y}}^{-1}(b) \cap f_{\mathsf{X}}^{-1}(a) \text{ otherwise} \end{cases} \tag{56}$$

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

$$\begin{aligned}
(\mathbf{L}'\mathsf{Y})^b_a &= \sum_{\omega\in\Omega} \mathbf{L}'^{\omega}_a \mathsf{Y}^b_\omega && (57)\\
&= \sum_{\omega\in f_{\mathsf{Y}}^{-1}(b)} \mathbf{L}'^{\omega}_a && (58)\\
&= \begin{cases} 0 & f_{\mathsf{Y}}^{-1}(b) \cap f_{\mathsf{X}}^{-1}(a) = \emptyset \\ \mathbf{K}'^{b}_a & \text{otherwise} \end{cases} && (59)\\
\implies (\mathbf{L}'\mathsf{Y}) &= \mathbf{K}' && (60)
\end{aligned}$$

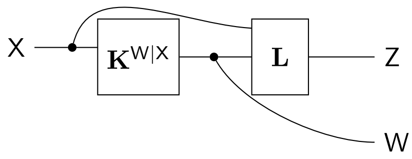
□

Definition 2.16 (Pushforward model)**.** For any variables $\mathsf{X} : \Omega \to A$, $\mathsf{Y} : \Omega \to B$ and any sample space model $\mathbb{K} : \mathsf{X} \to \mathrm{Id}_\Omega$, 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 2.10. 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 2.17 (Submodel)**.** Given $\mathbb{K} : \mathsf{X} \to \mathrm{Id}_\Omega$ and $\mathbb{L} : \mathsf{W}, \mathsf{X} \to \mathsf{Z}$, $\mathbb{L}$ is a submodel of $\mathbb{K}$ if

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



(61)

$$(\mathbb{K}^{\mathsf{Z},\mathsf{W}|\mathsf{Y}})^{w,z}_x = (\mathbb{K}^{\mathsf{W}|\mathsf{Y}})^w_x \mathbb{L}^z_{w,x} \tag{62}$$

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

Lemma 2.18 (Submodel existence)**.** *For any model $\mathbb{K} : \mathsf{X} \to \mathrm{Id}_\Omega$ (where Ω is a finite set), W and Y, there exists a submodel $\mathbb{L} : (\mathsf{W}, \mathsf{X}) \to \mathsf{Y}$.*

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

$$\mathbf{L}'^{y}_{wx} = \frac{\mathbb{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w, y|x)}{\mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x)} \qquad \forall x, w : \text{ the denominator is positive} \tag{63}$$

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

$$\begin{aligned}
\mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x)\mathbf{L}'^{y}_{xw} &= \mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x)\frac{\mathbb{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w,y|x)}{\mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x)} && \text{if } \mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x) > 0 && (64)\\
&= \mathbb{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w,y|x) && \text{if } \mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x) > 0 && (65)\\
&= 0 && \text{otherwise} && (66)\\
&= \mathbb{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w,y|x) && \text{otherwise} && (67)
\end{aligned}$$

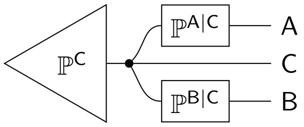
□

2.8 Conditional independence

We define conditional independence in the following manner:

For a *probability model* $\mathbb{P} : \mathsf{I} \rightharpoonup \mathrm{Id}_{\Omega}$ 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

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


$\mathbb{P}^{\mathsf{C}}$
$\mathbb{P}^{\mathsf{A}|\mathsf{C}}$
$\mathbb{P}^{\mathsf{B}|\mathsf{C}}$
A
C
B

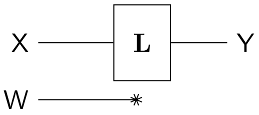
(68)

For an arbitrary model $\mathbf{N} : \mathsf{X} \rightharpoonup \mathrm{Id}_{\Omega}$ where $\mathsf{X} : \Omega \rightharpoonup X$, 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 $\mathbb{O} : \mathsf{I} \rightharpoonup \mathsf{X}$ 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 inapplicable 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:

$$\mathbf{K} =$$



(69)

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?

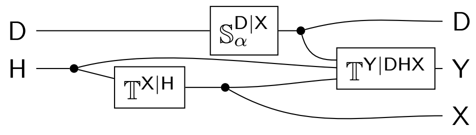
Suppose we will be given an observation, modelled by X taking values in X, and in response to this we can select any decision, modelled by D taking values in D. The process by which we choose a decision or mixture of decisions, is called a decision rule or a *strategy*, designated α and modelled by $\mathbf{S}_\alpha : \mathsf{X} \to \Delta(\mathsf{D})$[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 $\mathbf{P}_\alpha : \mathsf{I} \rightarrowtriangle (\mathsf{H}, \mathsf{X}, \mathsf{D}, \mathsf{Y})$. We assume that – holding the hypothesis fixed – the same observations are expected whatever strategy we choose: $\mathbb{P}_\alpha^{\mathsf{X|H}} = P_\beta^{\mathsf{X|H}}$ for all α, β. We assume that under each hypothesis, the decision chosen is always modelled by the the chosen strategy: $\mathbb{P}_\alpha^{\mathsf{D|HX}} = \mathbf{S}_\alpha \otimes \text{erase}_{\mathsf{H}}$. Finally, we assume that, holding the hypothesis fixed, the same consequences are expected under any strategy given the same observations and the same decision: $\mathbb{P}_\alpha^{\mathsf{Y|XHD}} = P_\beta^{\mathsf{Y|XHD}}$ for all α, β.

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

$$\mathbb{P}_\alpha =$$


D
H
$\mathbb{S}_\alpha^{\mathsf{D|X}}$
$\mathbb{T}^{\mathsf{X|H}}$
$\mathbb{T}^{\mathsf{Y|DHX}}$
D
Y
X

(70)

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

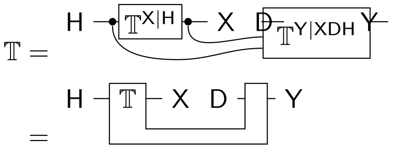
We will call the the see-do model along with the collection of strategies $\{\mathbb{T}^{\mathsf{XY|HD}}, \{\mathbb{S}_\alpha|\alpha \in \mathcal{A}\}\}$ a *standard decision problem.*

[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 70, 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.


$\mathbb{T} =$
H
$\mathbb{T}^{\mathsf{X}|\mathsf{H}}$
X
D
$\mathbb{T}^{\mathsf{Y}|\mathsf{XDH}}$
Y
=
H
$\mathbb{T}$
X
D
Y

(71)

(72)

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 $(\mathsf{D},\mathsf{H}) \rightarrowtriangle \mathsf{Y}$) 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 $\mathsf{X} : \Omega \rightarrowtriangle X$ and a statistical experiment $\{\mathbb{P}_\theta^{\mathsf{X}}\}_\Theta$, a decision set D and a loss $l : \Theta \times D \to \mathbb{R}$. A strategy $\mathbb{S}_\alpha^{\mathsf{D}|\mathsf{X}}$ is evaluated according to the risk functional $R(\theta, \alpha) := \sum_{x\in X} \sum_{d\in D} \mathbb{P}_\theta^{\mathsf{X}}(x) S_\alpha^{\mathsf{D}|\mathsf{X}}(d|x) l(h, d)$. A strategy $\mathbb{S}_\alpha^{\mathsf{D}|\mathsf{X}}$ is considered more desirable than $\mathbb{S}_\beta^{\mathsf{D}|\mathsf{X}}$ 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

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_\alpha}[\mathsf{Y}|h] &= \sum_{x\in X} \sum_{d\in D} \sum_{y\in Y} \mathbb{T}^{\mathsf{X}|\mathsf{H}}(x|h) \mathbb{S}_\alpha^{\mathsf{D}|\mathsf{X}}(d|x) \mathbb{T}^{\mathsf{Y}|\mathsf{HD}}(y|h, d) && (73)\\
&= \sum_{x\in X} \sum_{d\in D} \mathbb{T}^{\mathsf{X}|\mathsf{H}}(x|h) \mathbb{S}_\alpha^{\mathsf{D}|\mathsf{X}}(d|x) l(h, d) && (74)\\
&= R(h, \alpha) && (75)
\end{aligned}$$

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 70 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 $\mathcal{G}$:

$$\mathbb{P}_{\mathsf{X}=a}^{\mathsf{WXY}}(w, x, y) = \mathbb{P}^{\mathsf{W}}(w)\mathbb{P}^{\mathsf{Y}|\mathsf{XW}}(y|x, w)[\![x = a]\!] \tag{76}$$

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 $\mathbb{P}_{\mathsf{X}=a}^{\mathsf{WXY}}$, and $\mathbb{P}$ and $\mathbb{P}_{\mathsf{X}=a}$ are coupled by Equation 76. 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

$$\mathbb{P}^{\mathsf{V}_i} := \mathbb{U}^{\mathsf{V}_i|\mathsf{H}}(v|h) \text{ for some } h \in H \text{ and any } i \in A, v \in W \times X \times Y \tag{77}$$

$$\mathbb{P}_{\mathsf{X}_j=a}^{\mathsf{V}_j} := \mathbb{U}^{\mathsf{V}_j|\mathsf{HX}_j}(v|h, a) \text{ for some } h \in H \text{ and any } j \in B, v \in W \times X \times Y \tag{78}$$

We justify line 78 by noting that $\mathbb{U}^{\mathsf{V}_i|\mathsf{HX}}$ is a Markov kernel $H \times X \rightarrowtriangle W \times X \times Y$, which is the same type as the map $\mathbf{Q} := h, a \mapsto \mathbb{P}_{\mathsf{X}=a}$, and in addition Equation 76 ensures that defining $\mathbb{U}^{\mathsf{V}_i|\mathsf{HX}} := \mathbf{Q}$ is consistent via Lemma 2.7.

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 76 becomes

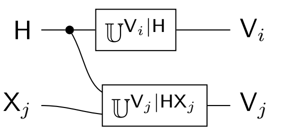
$$\mathbb{U}^{\mathsf{W}_j\mathsf{X}_j\mathsf{Y}_j|\mathsf{HX}_j}(w, x, y|h, a) = \mathbb{U}^{\mathsf{W}_i|\mathsf{H}}(w)\mathbb{U}^{\mathsf{Y}_i|\mathsf{X}_\mathsf{i}\mathsf{W}_\mathsf{i}\mathsf{H}}(y|x, w, h)[\![x = a]\!] \quad i \in A, j \in B \tag{79}$$

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

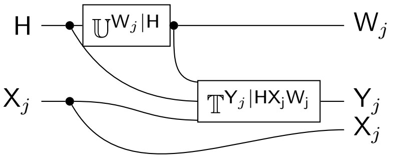
In most situations, A will be disjoint from B. While we don't necessarily want to rule out considering consequences to be equal to observations, we usually want to consider consequences that may take different values from observations.

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

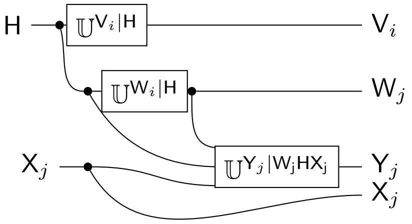
$$\mathbb{U}^{\mathsf{V}_i\mathsf{V}_j|\mathsf{HX}_j} = \tag{80}$$


H
$\mathbb{U}^{\mathsf{V}_i|\mathsf{H}}$
V_i
X_j
$\mathbb{U}^{\mathsf{V}_j|\mathsf{HX}_j}$
V_j

$$\mathbb{U}^{\mathsf{V}_j|\mathsf{HX}_j} = \tag{81}$$

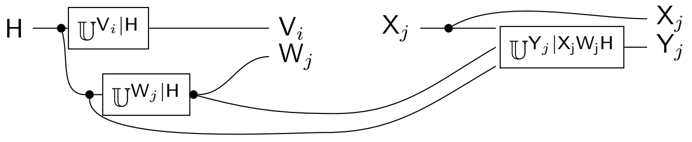

H
$\mathbb{U}^{\mathsf{W}_j|\mathsf{H}}$
W_j
X_j
$\mathbb{T}^{\mathsf{Y}_j|\mathsf{HX_jW_j}}$
Y_j
X_j

$$\implies \mathbb{U}^{\mathsf{V}_A\mathsf{V}_j|\mathsf{HX}_j} = \tag{82}$$


H
$\mathbb{U}^{\mathsf{V}_i|\mathsf{H}}$
V_i
$\mathbb{U}^{\mathsf{W}_i|\mathsf{H}}$
W_j
X_j
$\mathbb{U}^{\mathsf{Y}_j|\mathsf{W_jHX_j}}$
Y_j
X_j

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 82 defines a model $\mathbb{U}^{\cdot|\mathsf{HX}_j}$ which relates observations to consequences in the manner suggested by Equation 76. We will call $\mathbb{U}$ a "CBN model". We note that the model in Equation 82 looks like a 2-comb:

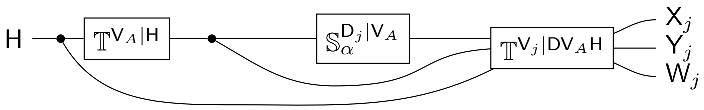
$$\tag{83}$$


H
$\mathbb{U}^{\mathsf{V}_i|\mathsf{H}}$
V_i
W_j
$\mathbb{U}^{\mathsf{W}_j|\mathsf{H}}$
X_j
X_j
$\mathbb{U}^{\mathsf{Y}_j|\mathsf{X_jW_jH}}$
Y_j

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

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

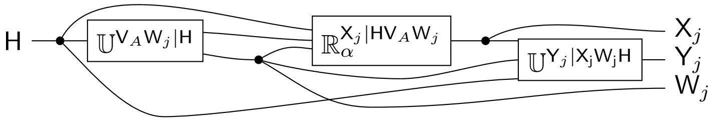
$\mathbb{P}_\alpha^{\mathsf{V}_B|\mathsf{H}} :=$



H
$\mathbb{T}^{\mathsf{V}_A|\mathsf{H}}$
$\mathbb{S}_\alpha^{\mathsf{D}_j|\mathsf{V}_A}$
$\mathbb{T}^{\mathsf{V}_j|\mathsf{D}\mathsf{V}_A\mathsf{H}}$
X_j
Y_j
W_j

(84)

$=$



H
$\mathbb{U}^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}}$
$\mathbb{R}_\alpha^{\mathsf{X}_j|\mathsf{H}\mathsf{V}_A\mathsf{W}_j}$
$\mathbb{U}^{\mathsf{Y}_j|\mathsf{X}_j\mathsf{W}_j\mathsf{H}}$
X_j
Y_j
W_j

(85)

$$=: \mathbb{Q}_\alpha^{\mathsf{V}_B|\mathsf{H}} \tag{86}$$

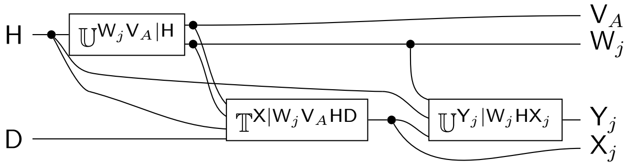
I think reusing the same H between U and T is a mistake here. Maybe not a big problem, but ideally one would check!

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 82, 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

$\mathbb{T}^{\mathsf{V}_A\mathsf{V}_j|\mathsf{H}\mathsf{D}} =$

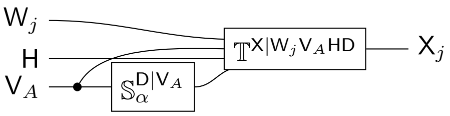


H
D
$\mathbb{U}^{\mathsf{W}_j\mathsf{V}_A|\mathsf{H}}$
$\mathbb{T}^{\mathsf{X}|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{D}}$
$\mathbb{U}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{H}\mathsf{X}_j}$
V_A
W_j
Y_j
X_j

(87)

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

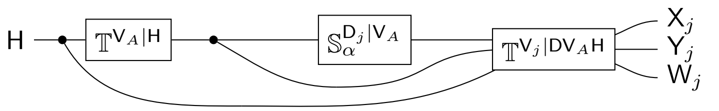
$\mathbb{R}_\alpha^{\mathsf{X}_j|\mathsf{V}_A\mathsf{W}_j\mathsf{H}} :=$



W_j
H
V_A
$\mathbb{S}_\alpha^{\mathsf{D}|\mathsf{V}_A}$
$\mathbb{T}^{\mathsf{X}|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{D}}$
X_j

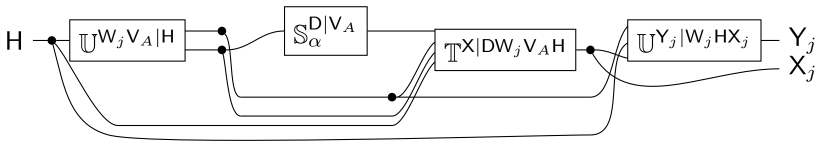
(88)

Then


H
$\mathbb{T}^{\mathsf{V}_A|\mathsf{H}}$
$\mathbb{S}_\alpha^{\mathsf{D}_j|\mathsf{V}_A}$
$\mathbb{T}^{\mathsf{V}_j|\mathsf{D}\mathsf{V}_A\mathsf{H}}$
X_j
Y_j
W_j

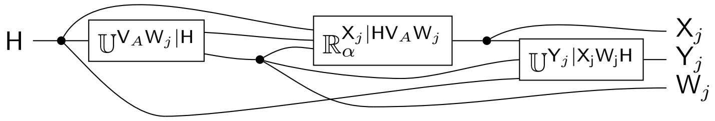
(89)

$=$


H
$\mathbb{U}^{\mathsf{W}_j\mathsf{V}_A|\mathsf{H}}$
$\mathbb{S}_\alpha^{\mathsf{D}|\mathsf{V}_A}$
$\mathbb{T}^{\mathsf{X}|\mathsf{D}\mathsf{W}_j\mathsf{V}_A\mathsf{H}}$
$\mathbb{U}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{H}\mathsf{X}_j}$
Y_j
X_j

(90)

$=$


H
$\mathbb{U}^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}}$
$\mathbb{R}_\alpha^{\mathsf{X}_j|\mathsf{H}\mathsf{V}_A\mathsf{W}_j}$
$\mathbb{U}^{\mathsf{Y}_j|\mathsf{X}_j\mathsf{W}_j\mathsf{H}}$
X_j
Y_j
W_j

(91)

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

$$\begin{aligned}\mathbb{P}_d^{\mathsf{W}_j|\mathsf{H}}(w|h) &= \mathbb{T}^{\mathsf{W}_j|\mathsf{HD}}(w_j|h,d) \qquad (92)\\ &\neq \mathbb{T}^{\mathsf{W}_j|\mathsf{HD}}(w_j|h,d') \qquad (93)\\ &= \mathbb{P}_{d'}^{\mathsf{W}_j|\mathsf{H}}(w|h) \qquad (94)\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}} &= \mathbb{T}^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}} \qquad (95)\\ &\neq \mathbb{U}^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}} \qquad (96)\\ &= \mathbb{Q}_\alpha^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}} \qquad (97)\end{aligned}$$

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

$$\begin{aligned}\mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x,d) &\neq \mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x,d') \qquad (98)\\ \text{and } \mathbb{T}^{\mathsf{X}_j\mathsf{W}_j\mathsf{V}_A|\mathsf{HD}}&(x,w,v|h,d) > 0 \qquad (99)\\ \text{and } \mathbb{T}^{\mathsf{X}_j\mathsf{W}_j\mathsf{V}_A|\mathsf{HD}}&(x,w,v|h,d') > 0 \qquad (100)\\ & \qquad (101)\end{aligned}$$

The latter conditions hold as if Equation 98 only held on sets of measure 0 then we could choose versions of the conditional probabilities such that the independence held.

Then

$$\begin{aligned}
\mathbb{P}_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) &= \mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x,d) && (102)\\
&\neq \mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x,d') && (103)\\
&= \mathbb{P}_{d'}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) && (104)\\
\implies \mathbb{P}_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) &\neq \mathbb{Q}_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) && (105)\\
\text{or } \mathbb{P}_{d'}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) &\neq \mathbb{Q}_{d'}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) && (106)
\end{aligned}$$

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 $y \in Y$

$$\begin{aligned}
\mathbb{P}_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) &= \mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j}(y|w,v,h,x) && (107)\\
&\neq \mathbb{U}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j}(y|w,v,h,x) && (108)\\
&\neq modelQ_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) && (109)
\end{aligned}$$

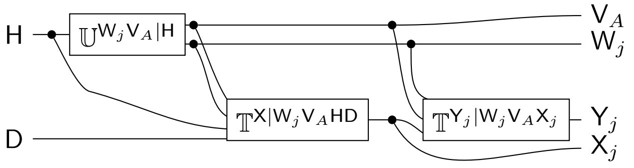
□

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

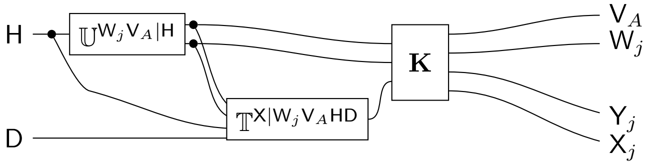
We say that $(\mathsf{V}_A, \mathsf{W}_j) \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$ expresses the notion that W_j is a *pre-choice variable* and $(\mathsf{W}_j, \mathsf{V}_A, \mathsf{X}_j)$ are *proxies for* D with respect to Y under conditions of full information. To justify this terminology,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

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


H
D
$\mathbb{U}^{\mathsf{W}_j\mathsf{V}_A|\mathsf{H}}$
$\mathbb{T}^{\mathsf{X}|\mathsf{W}_j\mathsf{V}_A\mathsf{HD}}$
$\mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{X}_j}$
V_A
W_j
Y_j
X_j

(110)

$=$


H
D
$\mathbb{U}^{\mathsf{W}_j\mathsf{V}_A|\mathsf{H}}$
$\mathbb{T}^{\mathsf{X}|\mathsf{W}_j\mathsf{V}_A\mathsf{HD}}$
$\mathbf{K}$
V_A
W_j
Y_j
X_j

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

(111)

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 2.6. □

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.*

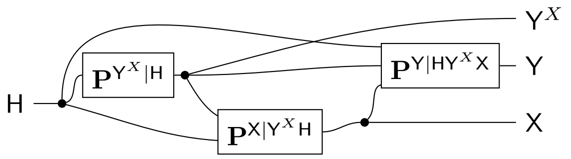
Definition 5.4 (What potential outcomes model: counterfactual extension)**.** Given observables X, Y and Y^X, Y^X are potential outcomes if they satisfy Definition 5.2 and for all $x \in X$, the individual potential outcome $\mathsf{Y}^x := \pi(x, \mathsf{Y})$ models the value Y would take if X took 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 2.18, we can represent the model $\mathbb{P}^{\mathsf{XYY}^X|\mathsf{H}}$ of these variables as

$\mathbb{P}^{\mathsf{XYY}^X|\mathsf{H}} =$



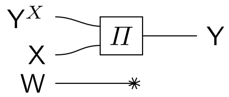
Y^X
H
$\mathbf{P}^{\mathsf{Y}^X|\mathsf{H}}$
$\mathbf{P}^{\mathsf{Y}|\mathsf{HY}^X\mathsf{X}}$
Y
$\mathbf{P}^{\mathsf{X}|\mathsf{Y}^X\mathsf{H}}$
X

(112)

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 112 somewhat. Firstly, $\mathbb{P}^{\mathsf{Y}|\mathsf{HY}^X\mathsf{X}}$ must always be represented a *selector kernel* $\Pi : X \times Y^{|X|} \rightarrow Y$, as shown by Lemma 5.5.

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

$\mathbf{Q} :=$



(113)

$$\iff \tag{114}$$

$$\mathbf{Q}^y_{(y^X,x,w)} = \Pi^y_{(x,y^X)} \qquad \forall y, y^X, x, w \tag{115}$$

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 113 and 115 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

$$\mathbb{T}^{\mathsf{Y}|\mathsf{HWXD}}(y|h,w,0,0) \neq \mathbb{T}^{\mathsf{Y}|\mathsf{HWXD}}(y|h,w,0,2) \tag{116}$$

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,

$$\mathbb{P}_0^{\mathsf{Y}|\mathsf{HD}}(y|h,0) = \sum_{x\in\{0,1\}, w\in W} \mathbb{T}^{\mathsf{W}|\mathsf{H}}(w|h)\mathbb{S}_0^{\mathsf{D}|\mathsf{W}}(0|w)\mathbb{T}^{\mathsf{X}|\mathsf{HWD}}(x|h,w,0)\mathbb{T}^{\mathsf{Y}|\mathsf{HWXD}}(y|h,w,x,0) \tag{117}$$

$$= \mathbb{T}^{\mathsf{Y}|\mathsf{HWXD}}(y|h,w,0,0) \tag{118}$$

$$\neq \mathbb{P}_2^{\mathsf{Y}|\mathsf{HD}} \tag{119}$$

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

$$\begin{aligned}\mathbb{P}_0^{\mathsf{Y}|\mathsf{HD}}(y|h,0) &= \sum_{y^X\in Y^2, x\in\{0,1\}} \mathbb{T}^{\mathsf{Y}^X|\mathsf{H}}(y^X|h)\mathbb{T}^{\mathsf{X}|\mathsf{HDY}^X}(x|h,0,y^X)\mathit{\Pi}(y|x,y^X) && (120)\\ &= \sum_{y^X}\mathbb{T}^{\mathsf{Y}^X|\mathsf{H}}(y^X|h)\mathit{\Pi}(y|0,y^X) && (121)\\ &= \sum_{y^X\in Y^2, x\in\{0,1\}} \mathbb{T}^{\mathsf{Y}^X|\mathsf{H}}(y^X|h)\mathbb{T}^{\mathsf{X}|\mathsf{HDY}^X}(x|h,2,y^X)\mathit{\Pi}(y|x,y^X) && (122)\\ &= \mathbb{P}_2^{\mathsf{Y}|\mathsf{HD}} && (123)\end{aligned}$$

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 119 and 123 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*

$$\mathsf{Y} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{WX} \tag{124}$$

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 124 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

$$\mathbb{P}[\mathsf{XDY}]^{ijk} = \mathbb{P}[\mathsf{X}]^i \mathbb{P}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{P}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \tag{125}$$

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

$$\mathbf{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^l_{kj} \tag{126}$$

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

$$\begin{aligned}
\mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj} &= \mathbf{T}\underline{\mathsf{X}}'^k_{hj} && (127)\\
&= \sum_l \mathbf{T}^{kl}_{hj} && (128)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^k && (129)\\
&= \mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj'} && (130)
\end{aligned}$$

Thus $\mathsf{X}' \perp\!\!\!\perp_{\mathbf{T}} \mathsf{D}'|\mathsf{H}'$ and so $\mathbf{T}[\mathsf{X}'|\mathsf{H}']$ exists (section 2.8) and $(\mathbf{T}, \mathsf{H}', \mathsf{D}', \mathsf{X}', \mathsf{Y}')$ is a see-do model.

Applying Equation 125 to $\mathbb{P}_{h,\mathbf{S}}$:

$$\begin{aligned}
\mathbb{P}_{h,\mathbf{S}}[\mathsf{XDY}]^{ijk} &= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} && (131)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} && (132)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]^j_i \mathbf{T}[\mathsf{X}'\mathsf{Y}'|\mathsf{H}'\mathsf{D}']^{ik}_{hj} && (133)\\
&= \mathbf{S}^j_i \mathbf{T}[\mathsf{X}'\mathsf{Y}'|\mathsf{H}'\mathsf{D}']^{ik}_{hj} && (134)\\
&= \mathbf{S}^j_i \mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^i_{hj} \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} && (135)\\
&= \mathbf{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbf{S}^j_i \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} && (136)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i &= \mathbb{P}_{h,\mathbf{Q}}[\mathsf{X}]^i && (137)\\
\mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} &= \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \text{ for some version of } \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}] && (138)
\end{aligned}$$

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

$$\mathbf{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbf{Q}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}]^l_{kj} \qquad (139)$$

Then by substitution of $\mathbf{Q}$ for $\mathbf{S}$ in Equation 131 and working through the same steps

$$\mathbb{P}_{h,\mathbf{S}}[\mathsf{XDY}]^{ijk} = \mathbf{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbf{Q}^j_i \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} \qquad (140)$$

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 $\mathbf{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]_{kj_1j_2\cdots j_{|W|}} = \delta[j_k]$.

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 $\mathbf{K}[\mathsf{X}_i|\mathsf{D}_i\mathsf{Z}\mathsf{Y}(D)]^{x_i}_{e_i z j_1 \ldots j_{|D|}} = \delta[j_{e_i}]^{x_i}$. 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 $y := (y_i)_{i\in|D|} \in Y^{|D|}$, $d := (d_i)_{i\in[n]} \in D^{[n]}$, $x := (x_i)_{i\in[n]} \in Y^{[n]}$.

$$\mathbf{K}[\mathsf{X}_n|\mathsf{D}_n\mathsf{Z}]^x_{dz} = \sum_{y\in Y^{|D|}} \mathbf{K}[\mathsf{X}_{[n]}|\mathsf{D}_n\mathsf{Z}\mathsf{Y}(D)]^x_{dzy}\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_{[n]}\mathsf{Z}]^y_{dz} \tag{141}$$

$$= \sum_{y\in Y^{|D|}} \prod_{i\in[n]} \delta[y_{d_i}]^{x_i}\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_n\mathsf{Z}]^y_{dz} \tag{142}$$

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 $\mathsf{Y}(S_d) = (\mathsf{Y}_d)_{d \in S_d}$.

$$\begin{aligned}
\mathbf{K}[\mathsf{X}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]^x_{dz} &= \sum_{y \in Y^{|D|}} \prod_{i \in Q_d} \delta[y_{d_i}]^{x_i} \mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_{[n]}\mathsf{Z}]^y_{dz} && (143)\\
&= \sum_{y' \in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}(R_d)\mathsf{Y}(R_d^C)|\mathsf{D}_{Q_d}\mathsf{D}_{Q_d^C}\mathsf{Z}]^{x_{Q_d} y'}_{d_{Q_d} d^C_{Q_d} z} && (144)\\
&= \sum_{y' \in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}_{R_d}\mathsf{Y}_{R_d^C}|\mathsf{D}_{Q_d}\mathsf{D}_{Q_d^C}\mathsf{Z}]^{x_{Q_d} y'}_{dz} && (145)\\
&= \sum_{y' \in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]^{x_{Q_d} y'}_{dz} \qquad \text{(using exchangeability)} && (146)
\end{aligned}$$

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

$$\begin{aligned}
\mathsf{I}_{K\cdot\cdot} &:= \mathsf{I}_K \setminus \mathsf{I}_L \setminus \mathsf{I}_J && (147)\\
\mathsf{I}_{KL\cdot} &:= \mathsf{I}_K \cap \mathsf{I}_L \setminus \mathsf{I}_J && (148)\\
\mathsf{I}_{K\cdot J} &:= \mathsf{I}_K \cap \mathsf{I}_J \setminus \mathsf{I}_L && (149)\\
\mathsf{I}_{KLJ} &:= \mathsf{I}_K \cap \mathsf{I}_L \cap \mathsf{I}_J && (150)\\
\mathsf{I}_{\cdot L\cdot} &:= \mathsf{I}_L \setminus \mathsf{I}_K \setminus \mathsf{I}_J && (151)\\
\mathsf{I}_{\cdot LJ} &:= \mathsf{I}_L \cap \mathsf{I}_J \setminus \mathsf{I}_K && (152)\\
\mathsf{I}_{\cdot\cdot J} &:= \mathsf{I}_J \setminus \mathsf{I}_K \setminus \mathsf{I}_L && (153)\\
\mathsf{O}_{K\cdot\cdot} &:= \mathsf{O}_K \setminus \mathsf{I}_N \setminus \mathsf{I}_J && (154)\\
\mathsf{O}_{KL\cdot} &:= \mathsf{O}_K \cap \mathsf{I}_L \setminus \mathsf{I}_J && (155)\\
\mathsf{O}_{K\cdot J} &:= \mathsf{O}_K \cap \mathsf{I}_J \setminus \mathsf{I}_L && (156)\\
\mathsf{O}_{KLJ} &:= \mathsf{O}_K \cap \mathsf{I}_L \cap \mathsf{I}_J && (157)\\
\mathsf{O}_{L\cdot} &:= \mathsf{O}_L \setminus \mathsf{I}_J && (158)\\
\mathsf{O}_{LJ} &:= \mathsf{O}_L \cap \mathsf{I}_J && (159)
\end{aligned}$$

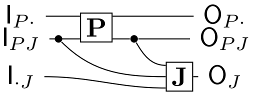
Also define

$$(\mathbf{P}, \mathsf{I}_P, \mathsf{O}_P) := \mathbf{K} \rightrightarrows \mathbf{L} \tag{160}$$
$$(\mathbf{Q}, \mathsf{I}_Q, \mathsf{O}_Q) := \mathbf{L} \rightrightarrows \mathbf{J} \tag{161}$$

Then

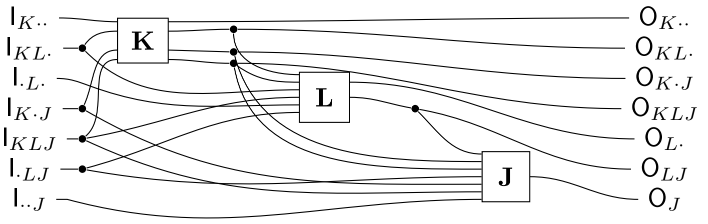
$$(\mathbf{K} \rightrightarrows \mathbf{L}) \rightrightarrows \mathbf{J} = \mathbf{P} \rightrightarrows \mathbf{J} \tag{162}$$

$=$



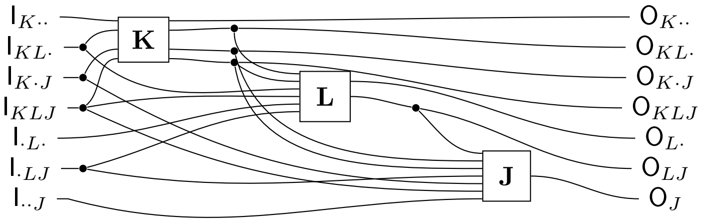
(163)

$=$



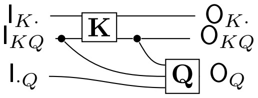
(164)

$\overset{perm}{=}$



(165)

$=$



(166)

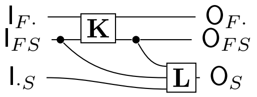
$$= \mathbf{K} \rightrightarrows (\mathbf{L} \rightrightarrows \mathbf{J}) \tag{167}$$

9 Appendix: String Diagram Examples

Recall the definition of *connection*:

Definition 9.1 (Connection)**.**

$$\mathbf{K} \rightrightarrows \mathbf{L} :=$$



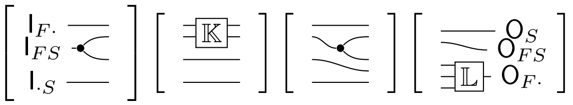
(168)

$$:= \mathbf{J} \tag{169}$$

$$\mathbf{J}_{yqr}^{zxw} = \mathbf{K}_{yq}^{zx}\mathbf{L}_{xqr}^{w} \tag{170}$$

Equation 168 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:

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



(171)

(172)

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 $(Y, \mathcal{Y})$

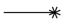
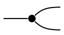
Note that this is a more general definition than the one used in the main paper; the version in the main paper is the restriction of this definition to finite sets.

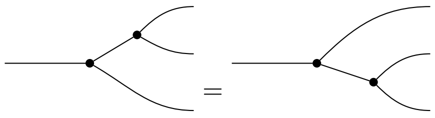
The *delta function* $\delta : X \to \Delta(\mathcal{X})$ is the Markov kernel defined by

$$\delta(x, A) = \begin{cases} 1 & x \in A \\ 0 & \text{otherwise} \end{cases} \tag{173}$$

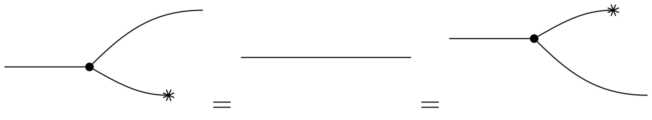
Fritz (2020) defines Markov categories in the following way:

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 $\text{copy}_X : X \to X \otimes X$ and a counit $\text{del}_X : X \to I$, depicted in string diagrams as

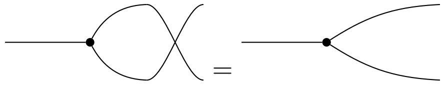
$\text{del}_X :=$  copy_X $:=$  (174) and satisfying the commutative comonoid equations



(175)

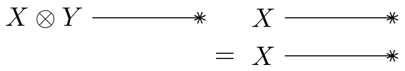


(176)

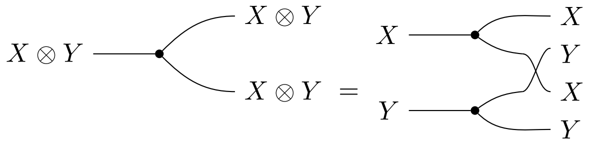


(177)

as well as compatibility with the monoidal structure

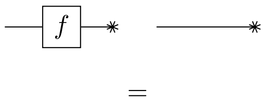


(178)



(179)

and the naturality of *del*, which means that



(180)

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

$$\mathbf{KL} : \mathsf{X} \to \mathit{\Delta}(\mathsf{Z}) \tag{181}$$

$$\text{defined by } (\mathbf{KL})(x, A) = \int_Y \mathbf{L}(y, A)\mathbf{K}(x, dy) \tag{182}$$

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

$$\mathbf{K} \otimes \mathbf{L} : (\mathsf{X}, \mathsf{W}) \to \mathit{\Delta}(\mathsf{Y}, \mathsf{Z}) \tag{183}$$

$$\text{defined by } \mathbf{K} \otimes \mathbf{L}(x, w, A \times B) = \mathbf{K}(x, A)\mathbf{L}(w, B) \tag{184}$$

The identity map is

$$\mathrm{Id}_{\mathsf{X}} : \mathsf{X} \to \mathit{\Delta}(\mathsf{X}) \tag{185}$$

$$\text{defined by}(\mathrm{Id}_X)(x, A) = \delta(x, A) \tag{186}$$

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

$$\mathrm{swap}_{\mathsf{X},\mathsf{Y}} : (\mathsf{X}, \mathsf{Y}) \to \mathit{\Delta}(\mathsf{Y}, \mathsf{X}) \tag{187}$$

$$\text{defined by}(\mathrm{swap}_{\mathsf{X},\mathsf{Y}})(x, y, A \times B) = \delta(x, B)\delta(y, A) \tag{188}$$

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

$$\mathrm{copy}_{\mathsf{X}} : \mathsf{X} \to \mathit{\Delta}(\mathsf{X}, \mathsf{X}) \tag{189}$$

$$\text{defined by}(\mathrm{copy}_X)(x, A \times B) = \delta_x(A)\delta_x(B) \tag{190}$$

and the erase map by

$$\mathrm{del}_{\mathsf{X}} : \mathsf{X} \to \mathit{\Delta}(*) \tag{191}$$

$$\text{defined by}(\mathrm{del}_X)(x, A) = \delta(*, A) \tag{192}$$

$$ \tag{193}$$

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 175 to 180 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`.

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: