Understanding Causal Primitives Using Modular Probability

David Johnston

October 18, 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.

To do this, we use a novel approach to probability modelling that we find is well suited to building causal mdoels. A typical approach to probability modelling is to construct a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ that serves as a top level model, along with a collection of random variables defined by measurable functions on this space, such that the particular quantities of interest can be obtained from conditional and marginal distributions on this space. Instead we consider a modelling context $\mathcal{M}$ that contains a collection of *probability components*, which are Markov kernels with named inputs and outputs. The names correspond to variables in the standard setting. Probability components with the right input and output types can be *connected*, an operation that yields a new probability component. We relate this back to the standard approach by equipping each probability component with a probability space and requiring that all components are the conditional probability distributions on their assigned spaces corresponding to their input and output labels.

Equipped with this foundation, we apply it to a variety of approaches to causal modelling, showing how it can enable understanding of different approaches in a common framework, and how it can represent assertions that were previously made "outside the model". First, we consider causal decision problems and derive *see-do models*, which reduce to statistical decision problems when augmented with the principle of expected utility. See-do models are a particular kind of probability component that we call a *comb*, which can be thought of as a probability model that needs something to be inserted into the middle. We consider causal graphical models, and show how under a very slight modification to the standard notation they induce see-do models, which allows us to formally connect *interventions* to *actions*. Finally, we consider potential outcomes models and show how we can formalise the typical assertion (which again, lives "outside the model") that potential outcomes represent counterfactual values. Potential outcomes models as typically used do not contain counterfactual assertions and in fact feature comb and insert components almost but not quite identical to combs and inserts found in causal graphical models.

I'm probably going to have to cut some of the above

2 Variables and Probability Models

2.1 Why are variables functions?

I don't think this subsection actually says much about what is coming next

Our intention is to clarify various "map-territory" issues related to causal inference. We start with a discussion of random variables as variables are often where we specify which real things our models are supposed to correspond to.

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 $\mathbb{P}$. A variable defined in this manner we will call a *formal variable*. In practice, however, it is very common to define variables with reference to:

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

For example, Pearl (2009) offers the same formal definiton as ours, but also says:

> 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.

Suppose we add the condition that a variable return the same value when given the same inquiry. Then we have an alternative definition of a variable:

- A set of ordered pairs (x, y), where x may take values of an attribute, measurement or inquiry and y may take values in a specified codomain Y
- Each x corresponds to exactly one y

This is almost a function, but it is not a function because a well-defined function requires a domain. It is what Menger (2003) calls a *qualitative statistical random variable*, and we will call a *vague variable* for short. Menger offers the examples of a vague variable that takes as an argument people in Chicago and returns their height in metres. Something that requires me to fly to Chicago with a tape measure in order to evaluate it is not a function.

However, some vague variables can be modeled by functions. You and I could start measuring people's heights and I tell you "X_i represents the ith person's height", then we go and measure someone's height and, in the end, we both decide decide that the appropriate value of X_1 is 1.78 metres. We agree because we each received an observation generated by the same measurement, and we mapped these observations to the same value. If we are sure we'll always agree on the result given observations from the same measurement, and we can model

all possible experiments as a set, then the relation between experiments and values that our respective "observation processes" induce is a function. In this sense X_1 is a well-defined vague variable.

Thus the definition of a vague variable X can sometimes be understood to be a statement like "let X be a formal variable that models the *measurement process* of interest". We say sometimes, because we do not claim to consider every usage to which random variables have ever been put.

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

$$\mathsf{X} = f_0(\xi_{\mathsf{X}}) \tag{1}$$
$$\mathsf{Y} = f_1(\mathsf{X}, \xi_{\mathsf{Y}}) \tag{2}$$

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 1 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):

$$\mathbb{P}_x^{\mathsf{XYZ}}(x', y, z) = \mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}(y|x, z)\mathbb{P}^{\mathsf{Z}}(z)[\![x = x']\!] \tag{3}$$

Equation 3 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 3 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 3, 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 3 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 3 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 3.

The condition of consistency allows us to check when non-standard products like Equation 3 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 3 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 9). 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 $\sum_{y \in Y} \mathbf{K}^y_x = 1$ for every $x \in X$. We will give Markov kernels the signature $\mathbf{K} : X \twoheadrightarrow Y$ 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:

$\mathbf{K} :=$

(4)

$\mathbf{P} :=$

(5)

Two Markov kernels $\mathbf{L} : X \nrightarrow Y$ and $\mathbf{M} : Y \nrightarrow Z$ have a product $\mathbf{LM} : X \nrightarrow Z$ 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:

$$\mathbf{LM} := \quad \text{-}\boxed{\mathbf{K}}\text{-}\boxed{\mathbf{M}}\text{-} \tag{6}$$

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

$$\mathbf{K} \otimes \mathbf{L} := \begin{array}{l} W \text{-}\boxed{\mathbf{K}}\text{-} Y \\ X \text{-}\boxed{\mathbf{L}}\text{-} Z \end{array} \tag{7}$$

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

$$\mathrm{id}_X := \quad X \text{-} X \tag{8}$$

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 $\mathrm{del}_X : X \nrightarrow \{*\}$ 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:

$$\mathrm{del}_X := \quad \text{-}\!\!*\, X \tag{9}$$

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

$$\mathrm{copy}_X := \quad X \text{-}\!\!\prec \begin{array}{l} X \\ X \end{array} \tag{10}$$

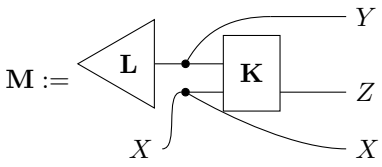
The swap map $\mathrm{swap}_{X,Y} : X \times Y \nrightarrow Y \times X$ defined by $(\mathrm{swap}_{X,Y})^{y',x'}_{x,y} = [\![x = x']\!][\![y = y']\!]$ swaps two inputs, and is represented by crossing wires:

$$\mathrm{swap}_X := \quad (11)$$

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

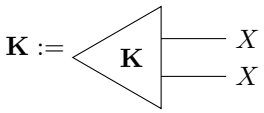
The Markov kernels introduced in the previous section can be though of as "conditional probability distributions without variables". We can use these to represent an operation very similar to Equation 3. Note that $P^{\mathsf{Y}|\mathsf{XZ}}$ must be represented by a Markov kernel $\mathbf{K} : X \times Z \nrightarrow Y$ and $\mathbb{P}^{\mathsf{Z}}$ by a Markov kernel $\mathbf{L} \in \Delta(Z)$. Then we can define a Markov kernel $\mathbf{M} : X \nrightarrow X \times Z$ representing $x \mapsto \mathbb{P}^{\mathsf{YZ}}_x(y, z)$ by



(12)

There is, however, a key difference between Equation 12 and Equation 3: 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 $\mathbf{K} : \{*\} \nrightarrow \Delta(X \times X)$. We could draw this:



(13)

We label both wires with the set X. However, say $X = \{0, 1\}$. Then $\mathbf{K}$ could be the kernel $\mathbf{K}^{x_1,x_2} = [\![x_1 = 0]\!][\![x_2 = 1]\!]$. In this case, both of its outputs must represent *different* variables, despite taking values in the same set. On the other hand, if $\mathbf{K}^{x_1,x_2} = 0.5[\![x_1 = x_2]\!]$ 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 $\mathsf{X} : \Omega \to\!\!\!\!\bullet\ A$ as a variable, where $\mathsf{X}^a_x = [\![a = f_{\mathsf{X}}(x)]\!]$.

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

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

The *unit* variable is the erase map $\mathsf{I} := \mathrm{del}_\Omega$, 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 $\mathbf{K} : \mathsf{X} \to\!\!\!\!\bullet\ \mathsf{Y}$ is a triple $(\mathbf{K}', \mathsf{X}, \mathsf{Y})$ where $\mathbf{K}' : A \to\!\!\!\!\bullet\ B$ is the *underlying kernel*, X is the *input index* and Y is the *output index*.

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

$$\mathbf{K} := \begin{array}{l}\mathsf{A}_1 \\ \mathsf{A}_2\end{array} \boxed{\mathbf{K}} \begin{array}{l}\mathsf{B}_1 \\ \mathsf{B}_2\end{array} \tag{14}$$

or

$$\mathbf{K} = \ (\mathsf{A}_1, \mathsf{A}_2) \boxed{\mathbf{K}[\mathbb{L}]} (\mathsf{B}_1, \mathsf{B}_2) \tag{15}$$

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

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

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

The category of indexed Markov kernels as morphisms and variables as objects is a Markov category (Appendix 9), and so a valid derivation based on the string diagram language for Markov categories corresponds to a valid theorem in this category. However, most of the diagrams we can form are not viable

candidates for models of our variables. For example, if X takes values in $\{0, 1\}$ we can propose an indexed Markov kernel $\mathbf{K} : \mathsf{X} \nrightarrow \mathsf{X}$ with $\mathbf{K}'^{b}_{a} = 0.5$ for all a, b. However, this is not a useful model of the variable X – it expresses something like "if we know the value of X, then we belive that X could take any value with equal probability".

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

Definition 2.3 (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:

$$f_{\mathsf{X}}^{-1}(a) \cap f_{\mathsf{Y}}^{-1}(b) = \emptyset \implies \left(\mathbf{K}'^{b}_{a} = 0\right) \vee \left(f_{\mathsf{X}}^{-1}(a) = \emptyset\right) \tag{16}$$

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

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

We take the idea of marginal distributions as fundamental.

Definition 2.4 (Marginal distribution)**.** Given a model $\mathbb{K} : \mathsf{X} \nrightarrow (\mathsf{Y}, \mathsf{Z})$, the marginal distribution of Y, written $\mathbb{K}^{\mathsf{Y}|\mathsf{X}}$, is obtained by marginalising over Z:

$$\mathbb{K}^{\mathsf{Y}|\mathsf{X}} := \quad \mathsf{X} \longrightarrow \boxed{\mathbf{K}'} \begin{array}{l} \longrightarrow \mathsf{Y} \\ \longrightarrow * \end{array} \tag{17}$$

$$\iff \tag{18}$$

$$(\mathbb{K}^{\mathsf{Y}|\mathsf{X}})^{y}_{x} = \sum_{z \in Z} \mathbf{K}'^{yz}_{x} \tag{19}$$

Definition 2.5 (Disintegration)**.** Given a model $\mathbb{K} : \mathsf{X} \nrightarrow (\mathsf{Y}, \mathsf{Z})$, a disintegration $\mathbb{L} : (\mathsf{X}, \mathsf{Y}) \nrightarrow \mathsf{Z}\ \mathsf{Y}$, written $\mathbb{K}^{\mathsf{Y}|\mathsf{X}}$, is obtained by marginalising over Z

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

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

$$\mathbf{K} = \quad \mathsf{X} \text{ —— } \boxed{\mathbb{L}} \text{ —— } \begin{matrix}\mathsf{Y}\\ \mathsf{Y}\\ \mathsf{Z}\end{matrix} \tag{20}$$

$$\iff \tag{21}$$

$$\mathbf{K}'^{y,y',z}_{x} = [\![y = y']\!]\mathbf{L}'^{y,z}_{x} \tag{22}$$

$$\tag{23}$$

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

$$(\mathsf{X}, \mathsf{Y}, \mathsf{Y}, \mathsf{Z})^{x,y,y',z}_{\omega} = [\![f_{\mathsf{Y}}(\omega) = y]\!][\![f_{\mathsf{Y}}(\omega) = y']\!](\mathsf{X}, \mathsf{Z})^{x,z}_{\omega} \tag{24}$$

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

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

$$\mathbf{K}'^{y,y',z}_{x} = [\![y = y']\!]\mathbf{L}'^{y,z}_{x} \tag{26}$$

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

$$[\![y = y']\!]\mathbb{L}'^{y,z}_{x} > 0 \implies y = y' \wedge \mathbb{L}'^{y,z}_{x} > 0 \tag{27}$$

$$\implies \left(f^{-1}_{\mathsf{X}}(x) = \emptyset\right) \vee \left(f^{-1}_{\mathsf{X}}(x) \cap f^{-1}_{\mathsf{Y}}(y) \cap f^{-1}_{\mathsf{Y}}(y) \cap f^{-1}_{\mathsf{Z}}(z) \neq \emptyset\right) \tag{28}$$

$$\tag{29}$$

□

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

$$\mathbf{K} = \quad \begin{matrix}\mathsf{X}\\ \mathsf{Y}\end{matrix} \text{ —— } \boxed{\mathbb{L}} \text{ —— } \begin{matrix}\mathsf{X}\\ \mathsf{Z}\end{matrix} \tag{30}$$

$$\iff \tag{31}$$

$$\mathbf{K}'^{x',z}_{x,y} = [\![x = x']\!]\mathbf{L}'^{z}_{x,y} \tag{32}$$

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

$$(\mathsf{X},\mathsf{Y},\mathsf{Y},\mathsf{Z})^{x,y,y',z}_{\omega} = [\![f_{\mathsf{Y}}(\omega) = y]\!][\![f_{\mathsf{Y}}(\omega) = y']\!](\mathsf{X},\mathsf{Z})^{x,z}_{\omega} \tag{33}$$

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

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

$$\mathbf{K}'^{x',z}_{x,y} = [\![x = x']\!]\mathbf{L}'^{z}_{x,y} \tag{35}$$

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

$$[\![x = x']\!]\mathbb{L}'^{z}_{x,y} > 0 \implies x = x' \wedge \mathbb{L}'^{z}_{x,y} > 0 \tag{36}$$

$$\implies \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) = \emptyset\right) \vee \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) \cap f_{\mathsf{Z}}^{-1}(z) \neq \emptyset\right) \tag{37}$$

$$\tag{38}$$

□

Consistency along with the notion of marginal distributions implies that, given some X and some $\mathbb{K} : \mathsf{Y} \twoheadrightarrow \mathrm{Id}_{\Omega}$, the pushforward $\mathbb{K}\mathbb{X}$ is the unique model $\mathsf{Y} \twoheadrightarrow \mathsf{X}$ 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 $\mathbb{X} : Id_{\Omega} \twoheadrightarrow \mathsf{X}$ given by $\mathbb{X} := (\mathsf{X}, Id_{\Omega}, \mathsf{X})$.*

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

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

$$\max_{\omega \in \Omega}(\mathrm{Id}_{\Omega},\mathsf{X})^{\omega',a}_{\omega} = \max_{\omega \in \Omega}[\![\omega = \omega']\!][\![\omega = f_{\mathsf{X}}(a)]\!] \tag{39}$$

$$= [\![\omega = f_{\mathsf{X}}(a)]\!] \tag{40}$$

$$= \mathbf{X}^{a}_{\omega} \tag{41}$$

Thus $\mathbb{X}$ satisfies consistency.

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

$$\max_{\omega \in \Omega}(\mathrm{Id}_{\Omega},\mathsf{X})^{\omega',a}_{\omega} = \max_{\omega \in \Omega}[\![\omega = \omega']\!][\![\omega = f_{\mathsf{X}}(b)]\!] \tag{42}$$

$$= [\![\omega = f_{\mathsf{X}}(b)]\!] \tag{43}$$

$$= 0 \tag{44}$$

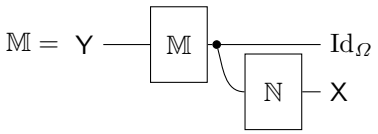
$$< \mathbb{K}^{b}_{\omega} \tag{45}$$

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

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

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

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



(46)

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

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

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

$$(\mathbb{K}\mathbb{X})^a_b = \sum_{\omega \in \Omega} \mathbb{P}^\omega_b \mathsf{X}^a_\omega \tag{47}$$

$$= \sum_{\omega \in \Omega} \mathbb{P}^\omega_b [\![a = f_{\mathsf{X}}(\omega)]\!] \tag{48}$$

$$= \sum_{\omega \in f^{-1}(a)} \mathbb{P}^\omega_b \tag{49}$$

□

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

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 $\mathsf{X} : \Omega \nrightarrow X$ and $\mathsf{Y} : \Omega \nrightarrow Y$ are variation independent, written $\mathsf{X} \perp_v \mathsf{Y}$, if for all $y \in f_{\mathsf{Y}}(\Omega) R(f_{\mathsf{Y}})$

$$f_{\mathsf{Y}}(\Omega) \times f_{\mathsf{X}}(\Omega) = \{(f_{\mathsf{Y}}(\omega), f_{\mathsf{X}}(\omega)) | \omega \in \Omega\} \tag{50}$$

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* $\mathbf{K} : \mathsf{X} \nrightarrow \mathsf{Y}$ *with* $\mathsf{X} : \Omega \nrightarrow X$ *and* $\mathsf{Y} : \Omega \nrightarrow Y$*, 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* $\mathbb{K} : \mathsf{X} \nrightarrow (\mathsf{Y}, \mathsf{Z})$*, if an indexed Markov kernel* $\mathbf{L} : (\mathsf{X}, \mathsf{Y}) \nrightarrow \mathsf{Z}$ *has the property* $\mathbf{K}_x'^{yz} = 0 \implies \mathbf{L}_{xy}'^{z} = 0$ *then* $\mathbf{L}$ *is also a model.*

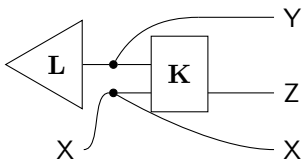
Proof. Because $\mathbb{K}$ is a model,

$$\begin{aligned}
\mathbf{L}_{xy}'^{z} > 0 &\implies \mathbf{K}_x'^{yz} > 0 && (51)\\
&\implies \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) \cap f_{\mathsf{Z}}^{-1}(z) \neq \emptyset\right) \vee \left(f_{\mathsf{X}}^{-1}(x) = \emptyset\right) && (52)\\
&\implies \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) \cap f_{\mathsf{Z}}^{-1}(z) \neq \emptyset\right) \vee \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) = \emptyset\right) && (53)
\end{aligned}$$

□

2.6 Truncated factorisation with variables

At this point, we can represent Equation 3 using models. Suppose $P^{\mathsf{Y}|\mathsf{XZ}}$ is an model $\mathbb{K} : (\mathsf{X}, \mathsf{Z}) \nrightarrow \mathsf{Y}$ and $\mathbb{P}^{\mathsf{Z}}$ an model $\mathbb{L} : \{*\} \nrightarrow \mathsf{Z}$. Then we can define an indexed Markov kernel $\mathbf{M} : \mathsf{X} \nrightarrow \mathsf{X}, \mathsf{Z}$ representing $x \mapsto \mathbb{P}_x^{\mathsf{YZ}}(y, z)$ by

$\mathbf{M} :=$ 

(54)

Equation 54 is almost identical to Equation 12, except it now specifies which variables each measure applies to, not just which sets they take values in. Like the original Equation 3, 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 54, 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 $\mathbf{K} : \mathsf{X} \twoheadrightarrow \mathrm{Id}_{\Omega}$. 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 $\mathbb{K} : \mathsf{X} \twoheadrightarrow \mathsf{Y}$, there is a sample space model $\mathbb{L} : \mathsf{X} \twoheadrightarrow Id_{\Omega}$ such that, defining $\mathbb{Y} := (\mathsf{Y}, Id_{\Omega}, \mathsf{Y})$, $\mathbb{L}\mathbb{Y} = \mathbb{K}$.*

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

$$\mathbf{L}'^{\omega}_{a} = \begin{cases} 0 & \text{if } f_{\mathsf{Y}}^{-1}(b) \cap f_{\mathsf{X}}^{-1}(a) = \emptyset \\ \mathbf{K}'^{b}_{a}[\![\omega = \omega_b]\!] & \text{for some } \omega_b \in f_{\mathsf{Y}}^{-1}(b) \text{ if } f_{\mathsf{X}}^{-1}(a) = \emptyset \\ \mathbf{K}'^{b}_{a}[\![\omega = \omega_{ab}]\!] & \text{for some } \omega_{ab} \in f_{\mathsf{Y}}^{-1}(b) \cap f_{\mathsf{X}}^{-1}(a) \text{ otherwise} \end{cases} \tag{55}$$

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

$$(\mathbf{L}'\mathsf{Y})^{b}_{a} = \sum_{\omega \in \Omega} \mathbf{L}'^{\omega}_{a} \mathsf{Y}^{b}_{\omega} \tag{56}$$

$$= \sum_{\omega \in f_{\mathsf{Y}}^{-1}(b)} \mathbf{L}'^{\omega}_{a} \tag{57}$$

$$= \begin{cases} 0 & f_{\mathsf{Y}}^{-1}(b) \cap f_{\mathsf{X}}^{-1}(a) = \emptyset \\ \mathbf{K}'^{b}_{a} & \text{otherwise} \end{cases} \tag{58}$$

$$\implies (\mathbf{L}'\mathsf{Y}) = \mathbf{K}' \tag{59}$$

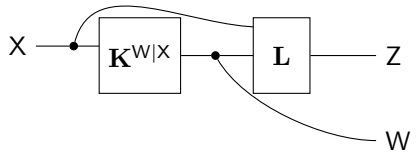
□

Definition 2.16 (Pushforward model)**.** For any variables $\mathsf{X} : \Omega \twoheadrightarrow A$, $\mathsf{Y} : \Omega \twoheadrightarrow B$ and any sample space model $\mathbb{K} : \mathsf{X} \twoheadrightarrow \mathrm{Id}_\Omega$, the pushforward $\mathbb{K}^{\mathsf{Y}|\mathsf{X}} := \mathbb{K}\mathbb{X}$ where $\mathbb{X} := (\mathsf{X}, \mathrm{Id}_\Omega, \mathsf{X})$.

The fact that the pushforward is a model is proved in Lemma 2.10. 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 $\mathbb{K} : \mathsf{X} \twoheadrightarrow \mathrm{Id}_\Omega$ and $\mathbb{L} : \mathsf{W}, \mathsf{X} \twoheadrightarrow \mathsf{Z}$, $\mathbb{L}$ is a submodel of $\mathbb{K}$ if

$$\mathbb{K}^{\mathsf{Z},\mathsf{W}|\mathsf{Y}} = \tag{60}$$



$$(\mathbb{K}^{\mathsf{Z},\mathsf{W}|\mathsf{Y}})^{w,z}_x = (\mathbb{K}^{\mathsf{W}|\mathsf{Y}})^w_x \mathbb{L}^z_{w,x} \tag{61}$$

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

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

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

$$\mathbf{L}'^y_{wx} = \frac{\mathbb{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w, y|x)}{\mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x)} \qquad \forall x, w : \text{ the denominator is positive} \tag{62}$$

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 $f^{-1}_{\mathsf{W}}(w) \cap f^{-1}_{\mathsf{X}}(x) = \emptyset$ or we can choose some $\omega_{wx} \in f^{-1}_{\mathsf{W}}(w) \cap f^{-1}_{\mathsf{X}}(x)$ 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

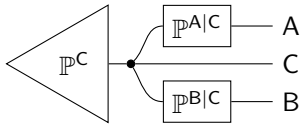
$$\begin{aligned}
\mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x)\mathbf{L}'^y_{xw} &= \mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x)\frac{\mathbb{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w, y|x)}{\mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x)} && \text{if } \mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x) > 0 && (63)\\
&= \mathbb{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w, y|x) && \text{if } \mathbb{K}^{\mathsf{W}|\mathsf{X}}(w|x) > 0 && (64)\\
&= 0 && \text{otherwise} && (65)\\
&= \mathbb{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w, y|x) && \text{otherwise} && (66)
\end{aligned}$$

□

2.8 Conditional independence

We define conditional independence in the following manner:

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

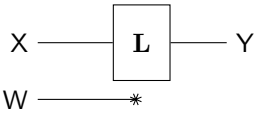
$\mathbf{P}^{\mathsf{ABC}} =$ 

(67)

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

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

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

$\mathbf{K} =$ 

(68)

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?

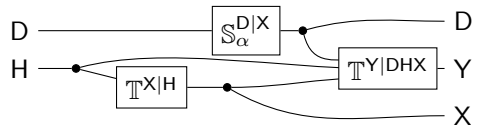
Suppose we will be given an observation, modelled by X taking values in X, and in response to this we can select any decision, modelled by D taking values in D. The process by which we choose a decision or mixture of decisions, is called a decision rule or a *strategy*, designated α and modelled by $\mathbf{S}_\alpha : \mathsf{X} \to \Delta(\mathsf{D})$[2]. We 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*.

[2]Recalling our discussion of variables, the strategy we ultimately choose could also be considered a vague variable. After we have made our choice, there is some "measurement process" by which we can determine which strategy we chose and which will always yield a consistent output given the same act of choosing. The reason we don't do so is that it would necessitate an extension of our theory to uncountable sets, which complicates the story.

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 $\mathbf{P}_\alpha : \mathsf{I} \to (\mathsf{H}, \mathsf{X}, \mathsf{D}, \mathsf{Y})$. We assume that – holding the hypothesis fixed – the same observations are expected whatever strategy we choose: $\mathbb{P}_\alpha^{\mathsf{X}|\mathsf{H}} = P_\beta^{\mathsf{X}|\mathsf{H}}$ for all α, β. We assume that under each hypothesis, the decision chosen is always modelled by the the chosen strategy: $\mathbb{P}_\alpha^{\mathsf{D}|\mathsf{H}\mathsf{X}} = \mathbf{S}_\alpha \otimes \mathrm{erase}_{\mathsf{H}}$. Finally, we assume that, holding the hypothesis fixed, the same consequences are expected under any strategy given the same observations and the same decision: $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{H}\mathsf{D}} = P_\beta^{\mathsf{Y}|\mathsf{H}\mathsf{D}}$ for all α, β.

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

$\mathbb{P}_\alpha =$ 

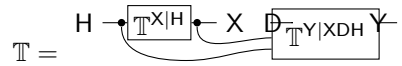
(69)

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

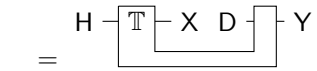
3.1 Combs

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 69, 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.

$\mathbb{T} =$ 

(70)

$=$ 

(71)

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 $(\mathsf{D}, \mathsf{H}) \to \mathsf{Y}$) 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 strate-

gies 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 $\mathsf{X} : \Omega \twoheadrightarrow X$ 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

$$\mathbb{E}_{\mathbb{P}_\alpha}[\mathsf{Y}|h] = \sum_{x\in X}\sum_{d\in D}\sum_{y\in Y} \mathbb{T}^{\mathsf{X}|\mathsf{H}}(x|h)\mathbb{S}_\alpha^{\mathsf{D}|\mathsf{X}}(d|x)\mathbb{T}^{\mathsf{Y}|\mathsf{HD}}(y|h,d) \tag{72}$$

$$= \sum_{x\in X}\sum_{d\in D} \mathbb{T}^{\mathsf{X}|\mathsf{H}}(x|h)\mathbb{S}_\alpha^{\mathsf{D}|\mathsf{X}}(d|x)l(h,d) \tag{73}$$

$$= R(h,\alpha) \tag{74}$$

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 69 is the risk of that strategy under hypothesis h.

4 Causal Bayesian Networks

Vague variable = observable variable. The connotations of the latter make explanations a bit more intuitive. Also, somehow explain "latent variables are things we use to make models"

What is a causal Bayesian network? We will consider a simplified case where a single node may be intervened on. With this condition, according to Pearl (2009), a causal Bayesian network is a probability model $\mathbb{P}$, a collection of interventional probability models $\{\mathbb{P}_{\mathsf{X}=a}|a \in X_i\}$ and a directed acyclic graph $\mathcal{G}$ whose nodes are identified with some collection of variables, which we can group into three variables $\{\mathsf{W},\mathsf{X},\mathsf{Y}\}$, where W is the sequence of variables associated with the parents of X in $\mathcal{G}$, X is the "intervenable" node of $\mathcal{G}$ and Y are associated with the other nodes. The interventional probability models must all obey the truncated factorisation condition with respect to $\mathcal{G}$:

$$\mathbb{P}^{\mathsf{WXY}}_{\mathsf{X}=a}(w,x,y) = \mathbb{P}^{\mathsf{W}}(w)\mathbb{P}^{\mathsf{Y}|\mathsf{XW}}(y|x,w)[\![x=a]\!] \tag{75}$$

Prove this is equivalent to the normal definition

What are these variables $\{\mathsf{W},\mathsf{X},\mathsf{Y}\}$, and what do we mean they are distributed according to $\mathbb{P}$? To begin with, it means that observations are modeled by a sequence of variables $\mathsf{V}_A := (\mathsf{W}_i,\mathsf{X}_i,\mathsf{Y}_i)_{i\in A}$, for which we assume the triples $(\mathsf{W}_i,\mathsf{X}_i,\mathsf{Y}_i)$ are mutually independent and identically distributed but we are not sure exactly how they are distributed. This can be captured by introducing a latent variable Hrepresenting the distribution of $\mathsf{V}_i := (\mathsf{W}_i,\mathsf{X}_i,\mathsf{Y}_i)$ for any $i \in A$[3]. Then we define a model of observations such that $\mathbb{T}^{\mathsf{V}_i|\mathsf{H}} = \mathbb{T}^{\mathsf{V}_j|\mathsf{H}}$ (for any $i,j \in A$) and $\mathsf{V}_i \perp\!\!\!\perp_{\mathbb{T}} \mathsf{V}_j|\mathsf{H}$. Then, the assumption "given a probability model $\mathbb{P}$" can be identified with the definition

$$\mathbb{P} := \mathbb{T}^{\mathsf{V}_i|\mathsf{H}}(\cdot|h) \text{ for some } h \in H \text{ and any } i \in A \tag{76}$$

[3]Under the theory introduced we are implicitly assuming H to be finite. It would clearly be desirable to extend the theory so that we can weaken this assumption, but it doesn't prevent us from explaining the basic idea.

What do the interventional probability models represent? We have already established on the basis of observations that the variables $\mathsf{W}, \mathsf{X}, \mathsf{Y}$ don't represent "observables" in the sense we discuss in Section 2.1 – we cannot explain which observation specifically W represents. We will suppose, as with observations, that there is some set B such that V_B are a sequence of observables modeled by the interventional models but we will leave B unspecified for now.

We also know that interventional probability models are coupled to observational models via H. That is Equations 75 and 76 together indicate that for different values $h, h' \in H$, we will generally get different sets of interventional models. We can define a collection of models of interventions $\mathbb{U}^{\mathsf{V}_B|\mathsf{H}}_{\mathsf{X}_B=a}$ such that V_i are independent and identically distributed conditional on H and for any $i \in B$, $j \in A$:

$$\mathbb{U}^{\mathsf{W}_i\mathsf{X}_i\mathsf{Y}_i|\mathsf{H}}_{\mathsf{X}_i=a}(w, x, y|h) = \mathbb{T}^{\mathsf{W}_j|\mathsf{H}}(w)\mathbb{T}^{\mathsf{Y}_j|\mathsf{X}_j\mathsf{W}_j\mathsf{H}}(y|x, w, h)[\![x = a]\!] \tag{77}$$

This is just the same as Equation 75, except we make the coupling between $\mathbb{T}$ and $\mathbb{U}$ via H explicit. We will also modify the notation in one more way. Instead of considering a collection of interventional models, we'll consider the map $\mathbf{U} := a \mapsto \mathbb{U}^{\mathsf{V}_B|\mathsf{H}}_{\mathsf{X}_B=a}$. $\mathbf{U}$ is a Markov kernel, and to make it a model we need to specify the domain and codomain indices. It can inherit the codomain index (V_B) from the original interventional interventional models, and its domain index is clearly going to be (H, RVQ) for some Q – i.e. it will also inherit H in its domain index.

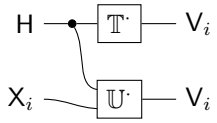
To work out what the remaining variable Q ought to be, the question we need to answer is: when we supply a value a to $\mathbf{U}$, *which observable variable are we saying takes the value* a? Consider that we refer to interventions by invoking the variable X_i, as in the subscript "$\mathsf{X}_i = a$" (or, in alternative notation, $do(\mathsf{X}_i = a)$). Furthermore, Equations 75 and 77 force X to be deterministically equal to the argument of U. For these reasons (and others we will see below), we think it is reasonable to choose X_B to be the remaining domain index of $\mathbb{U}$. Thus we have:

$$\mathbb{U}^{\mathsf{V}_B|\mathsf{H}\mathsf{X}_B}(v|h, a) = \mathbb{U}^{\mathsf{V}_B|\mathsf{H}}_{\mathsf{X}_B=a}(v|h) \tag{78}$$

$$\mathbb{U}^{\mathsf{W}_i\mathsf{X}_i\mathsf{Y}_i|\mathsf{H}\mathsf{X}_i}(w, x, y|h, a) = \mathbb{U}^{\mathsf{W}_i\mathsf{X}_i\mathsf{Y}_i|\mathsf{H}}_{\mathsf{X}_i=a}(w, x, y|h) \tag{79}$$

And we stipulate that $(\mathsf{W}_i\mathsf{X}_i\mathsf{Y}_i) \perp\!\!\!\perp_{\mathbb{U}} \mathsf{V}_{B\setminus\{i\}}|\mathsf{H}\mathsf{X}_i$, which is makes the second line well-defined (without this we would have to condition on X_B).

We intentionally left the sets A and B vague. Let's consider the possibility that they overlap; that is, there is some $i \in A \cap B$. We can express the coupling between observational and interventional models by joining them with a copymap, as follows:



(80)

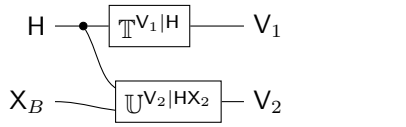
However, 80 cannot usually be a model. Because the variable V_i appears twice, unless we force the output at each terminal to be deterministically equal,

which limits us to modelling deterministic functions of H, then this model is inconsistent (see Lemma 2.6). Thus we can conclude that, usually, $B \cap A = \emptyset$, that is to say observations are typically distinct from consequences.

Purely to avoid introducing new graphical notation, we will suppose that $A = \{1\}$ and $B = \{2\}$. Because the sequences are conditionally independent and identically distributed, we can construct a model of arbitrary length sequences A and B from length-one examples. However, we will leave this construction implicit.

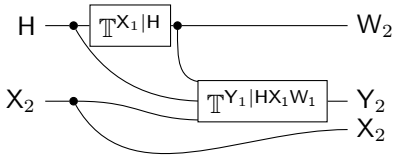
If we assume that the triples are $(\mathsf{W}_i, \mathsf{X}_i, \mathsf{Y}_i)$ are mutually variationally independent for all $i \in A \cup B$, then we can also extend $\mathbb{T}$ as follows:

$$\mathbb{T}^{\mathsf{V}_{A\cup B}|\mathsf{H}\mathsf{X}_B} :=$$



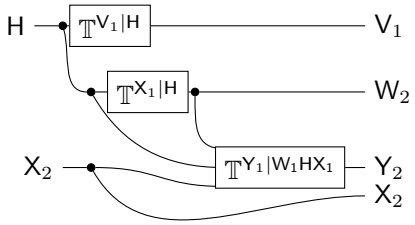
(81)

$$\mathbb{U}^{\mathsf{W}_i\mathsf{X}_i\mathsf{Y}_i|\mathsf{H}\mathsf{X}_i} =$$



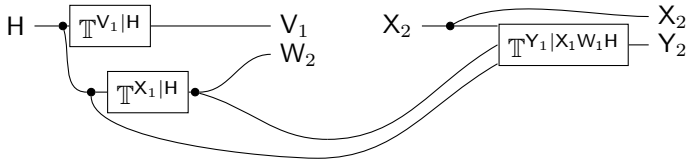
(82)

$$\implies \mathbb{T}^{\mathsf{V}_{A\cup B}|\mathsf{H}\mathsf{X}_B} =$$



(83)

So we have in Equation 83 a definition of a model $\mathbb{T}$ which, we argue, is the general form of a model of observables implied by the original definition of a causal Bayesian network (recall that our definition restricted interventions to target a single node). The model in Equation 83 looks like a 2-comb:



(84)

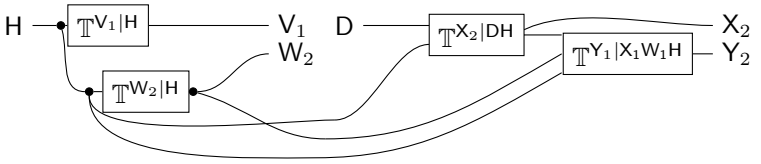
To actually *be* a 2-comb we require a family of "inserts" $\mathbb{S}^{\mathsf{X}_B|\mathsf{V}_A\mathsf{W}_B}$ that, in combination with $\mathbb{T}^{\mathsf{V}_{A\cup B}|\mathsf{H}\mathsf{X}_B}$, generate models $\mathbb{P}^{\mathsf{V}_{A\cup B}|\mathsf{H}}$. But still, we can make a reasonable guess guess: perhaps $\mathbb{T}^{\mathsf{V}_{A\cup B}|\mathsf{H}\mathsf{X}_B}$ is a see-do model, and the appropriate family of inserts are strategies.

We can weaken this somewhat, in that we don't strictly need X_B to be a variable representing decisions. This might be desirable, because the range of

X_j, $j \in B$ is exactly the same as the range of X_i, $i \in A$, but in general, there's no reason to expect that the observations we make will have the same range as the decisions available. To return to our favourite example, we might have observations X of body mass index, but the decision D we are considering might be whether or not to go on a diet. A more generic see-do model might then be of the form $\mathbb{T}^{\mathsf{V}_{A\cup B}|\mathsf{HD}}$ for some decision variable D.

Instead, we can introduce a see-do models $\mathbb{T}^{\mathsf{V}_{A\cup B}|\mathsf{HD}}$ with a generic variable D representign decisions. If and only if it exhibits the conditional independence $\mathsf{V}_B \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|(\mathsf{H}, \mathsf{X}_B)$, $\mathbb{T}^{\mathsf{V}_{A\cup B}|\mathsf{HX}_B}$ will be a conditional probability of $\mathsf{T}^{\mathsf{X}_B|\mathsf{W}_B\mathsf{HD}}$. In such a case, we can represent the see-do model

$\mathbb{T}^{\mathsf{V}_{A\cup B}|\mathsf{HD}} =$



(85)

Proof. (sketch) We require $\mathsf{V}_A \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$ (implied by see-do assumption), $\mathsf{W}_B \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$ (implied by $\mathsf{V}_B \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|(\mathsf{H}, \mathsf{X}_B)$ along with $\mathsf{W}_B \perp\!\!\!\perp_{\mathbb{T}} \mathsf{X}_B|\mathsf{H}$), $\mathsf{W}_B \perp\!\!\!\perp_{\mathbb{T}} \mathsf{X}_B|\mathsf{HD}$ (implied by $\mathsf{W}_B \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|(\mathsf{H}, \mathsf{X}_B)$ and $\mathsf{W}_B \perp\!\!\!\perp_{\mathbb{T}} \mathsf{X}_B|\mathsf{H}$), $\mathsf{Y}_B \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$ (assumed). □

In summary, when we have a causal Bayesian network where it is possible to intervene on a variable X, we can construct a see-do model $\mathbb{T}^{\mathsf{V}_A\mathsf{V}_B|\mathsf{HD}}$ with the conditional independence $\mathsf{V}_B \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|(\mathsf{H}, \mathsf{X}_B)$. This conditional idependence resembles the "limited invariance" condition proposed by Heckerman and Shachter (1995) as an account of causation.

The independence $\mathsf{Y}_B \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|(\mathsf{H}, \mathsf{X}_B)$ can be interpreted as expressing the property "if I knew H, then the effect of my decision D on Y_B is determined entirely by its effect on X_B". If this independence holds, then under conditions of full knowledge about the relationship betwen X and Y, X acts as a proxy for D in controlling Y. For short, we say X is a *full-knowledge proxy for* D. This assumption does not by itself permit us to reason about the effect of D on Y_B by separately considering the effect of D on X_B and the relationship between X_B and Y_B. For example, suppose H, D, Y_B and X_B are all binary, with D representing "do I go on a diet?", Y_B representing "do I experience heart disease?" and X_B an indicator for obesity based on my body mass index. Suppose that my model is

$$\mathbb{T}^{\mathsf{Y}_B\mathsf{X}_B|\mathsf{DH}}(y, x|d, h) = \begin{cases} 0.5[\![x = y]\!] & d \in \{0, 1\}, h = 0 \\ 0.5[\![x = d]\!] & d \in \{0, 1\}, h = 1 \end{cases} \tag{86}$$

We can verify that $\mathsf{Y}_B \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|(\mathsf{H}, \mathsf{X}_B)$. Under $h = 0$, if I am not obese I do not experience heart disease, but my diet has no effect. Under $h = 1$ if I diet I avoid obesity but obesity has no impact on my chance of heart disease. While a diet could reduce obesity and obesity could reduce heart disease, a diet can

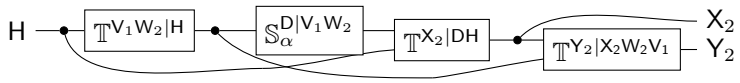
under no circumstances help me avoid heart disease. We can consider extending $\mathbb{T}$ with a prior that gave each hypothesis a probability of 0.5 and find that, for example, a diet reduces obesity in expectation and obesity reduces heart disease in expectation but a diet will not reduce heart disease.

However, there is an additional assumption that will allow "two part reasoning" of this type. That assumption is $\mathsf{V}_B \perp\!\!\!\perp \mathsf{H} | \mathsf{V}_A \mathsf{X}_B$. Explaining this assumption formally is beyond the theory presented here, but it can be thought of as expressing the assumption that the interventional map is precisely identifiable. That is, the observational data is enough to precisely identify the latent variable H up to the kernel of $\mathbb{T}^{\mathsf{V}_B|\mathsf{X}_B\mathsf{H}}$.

in the context of conditionally IID data, we can get "precise identifiability" with infinite data, but infinite data requires continuous measurable sets

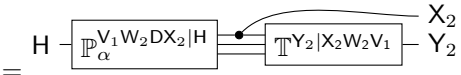
.

Given this additional assumption of "identifiability", we obtain the property $\mathsf{V}_B \perp\!\!\!\perp \mathsf{D} | \mathsf{V}_A$. If this holds, we can consider the application of an arbitrary strategy $(\mathsf{V}_A, \mathsf{W}_B) \to \mathsf{D}$ and write the resulting probabilistic model:



(87)

$=$



(88)

draw this diagram

As we can see in Equation 88, the probabilistic model decomposes into two parts - one that depends on the unknown H and the other that depends only on known quantities. The opportunity to reason in this manner ties in neatly with Pearl (2018) 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.

The assumption that smoking is a full-knowledge proxy for some action affecting cigarette advertising is an additional assumption, not a consequence of anything in the original causal Bayesian network. One might be inclined to say that an assumption of this nature is implicit when one chooses to try calculating *do*(smoking) to begin with. Perhaps so; here we are aiming to make such implicit assumptions explicit.

5 Potential outcomes with and without counterfactuals

Potential outcomes is a widely used approach to causal modelling characterised by its use of "potential outcome" random variables. Potential outcome random variables are typically noted for being given counterfactual interpretations. For example, suppose have something we want to model, call it TYT ("The Y Thing"), which we represent with a variable Y. Suppose we want to know how TYT behaves under different regimes 0 and 1 under which we want to know about TYT, and we use a variable W to indicate which regime holds at a given point in time. A potential outcomes model will introduce the two additional "potential outcome" variables $(\mathsf{Y}(0), \mathsf{Y}(1))$. What these variables represent can be given a counterfactual interpretation like "$\mathsf{Y}(0)$ represents what TYT would be under regime 0, whether or not regime 0 is the actual regime" and similarly "$\mathsf{Y}(1)$ represents what TYT would be under regime 1, whether or not regime 1 is the actual regime". Note that we say "what TYT would be" rather that "what Y would be" as "what would Y be if W was 0 if W was actually 1" is not a question we can ask of random variables, but it is one that might make sense for the things we use random variables to model.

This is a key point, so it is worth restating: the assumption that potential outcome variables agree with "the value TYT would take" under fixed regimes regardless of the "actual" value of the regime seems to be a critical assumption that distinguishes potential outcome variables from arbitrary random variables that happen to take values in the same space as Y. However, this assumption can only be stated by making reference to the informally defined "TYT" and the informal distinction between the supposed and the actual value of the regime.

The potential outcomes framework features other critical assumptions that relate potential outcome variables to things that are only informally defined. For example, Rubin (2005) defines the *Stable Unit Treatment Value Assumption* (SUTVA) as:

> SUTVA (stable unit treatment value assumption) [...] comprises two subassumptions. First, it assumes that there is no interference between units (Cox 1958); that is, neither $Y_i(1)$ nor $Y_i(0)$ is affected by what action any other unit received. Second, it assumes that there are no hidden versions of treatments; no matter how unit i received treatment 1, the outcome that would be observed would be $Y_i(1)$ and similarly for treatment 0

"Versions of treatments" do not appear within typical potential outcomes models, so this is also an assumption about how "the thing we are trying to model" behaves rather than an assumption stated within the model.

Given informal assumptions like this, one may be motivated to "formalize" them. More specifically, one might be motivated to ask whether there is some larger class of models that, under conditions corresponding to the informal conditions above yield regular potential outcome models?

I have a vague intuition here that you always need some kind of assumption like "my model is faithful to the real thing", but if you are stating fairly specific conditions in English you should also be able to state them mathematically. Among other reasons, this is useful because it's easier for other people to know what you mean when you state them.

The approach we have introduced here, motivated by decision problems, has in the past been considered a means of avoiding counterfactual statements, which has been considered a positive by some (Dawid, 2000) and a negative by others:

> [...] Dawid, in our opinion, incorrectly concludes that an approach to causal inference based on "decision analysis" and free of counterfactuals is completely satisfactory for addressing the problem of inference about the effects of causes.(Robins and Greenland, 2000)

It may be surprising to some, then, that we can use see-do models to formally state these key assumptions associated with potential outcomes models. Furthermore, we will argue that potential outcomes are typically a strategy to motivate inductive assumptions in see-do models, and we will show that the counterfactual interpretation is unnecessary for this purpose.

5.1 Potential outcomes in see-do models

A basic property of potential outcomes models is the relation between variables representing actual outcomes and variables representing potential outcomes, which was stated informally in the opening paragraph of this section.

In the following definition, $\mathsf{Y}(W) = (\mathsf{Y}(w))_{w \in W}$.

Definition 5.1 (Potential outcomes)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, a collection of variables $\{\mathsf{Y}, \mathsf{Y}(W), \mathsf{W}\}$ where Y and $\mathsf{Y}(W)$ are random variables and W could be either a state or a random variable is a *potential outcome submodel* if $\mathbf{K}[\mathsf{Y}|\mathsf{W}\mathsf{Y}(W)]$ exists and $\mathbf{K}[\mathsf{Y}|\mathsf{W}\mathsf{Y}(W)]_{ij_1j_2\cdots j_{|W|}} = \delta[j_i]$.

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

We allow X to be a state or a random variable to cover the cases where potential outcomes models feature as submodels of observation models (in which case X is a random variable) or as submodels of consequence models (in which case X may be a state variable).

As an aside that we could define stochastic potential outcomes if we allow the variables $\mathsf{Y}(x)$ to take values in $\Delta(Y)$ rather than in Y, and then require $\mathbf{K}[\mathsf{Y}|\mathsf{X}\mathsf{Y}(X)]_{ij_1j_2\dots j_{|X|}} = j_i$ (where j_i is an element of $\Delta(Y)$). This is more complex to work with and rarely seen in practice, but it is worth noting that Definition 5.1 can be generalised to cover models where $\mathsf{Y}(x)$ describes the value Y would take if X were x *with uncertainty*.

An arbitrary see-do model featuring potential outcome submodels does not necessarily allow for the formal statement of the counterfactual interpretation of potential outcomes. Here we use TYT ("the actual thing") and "regime" to refer to the things we are actually trying to model. We require that $\mathsf{Y} \overset{a.s.}{=} \mathsf{Y}(w)$ conditioned on $\mathsf{W} = w$. If we add an interpretation to this model saying Y represents TYT and W represents the regime, then we have "for all w, $\mathsf{Y}(w)$ is equal to Y which represents TYT under the regime w". However, this does not guarantee that our model has anything that reasonably represents "what TYT would be equal to under supposed regime w if the regime is actually w'".

We propose *parallel potential outcome submodels* as a means of formalising statements about what how TYT behaves under "supposed" and "actual" regimes:

Definition 5.2 (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 $\mathbf{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]_{kj_1j_2\dots j_{|W|}} = \delta[j_k]$.

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.

5.2 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 5.3 (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 5.4 (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 5.5 (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 5.6 (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 5.7 (Potential outcomes representation)**.** *Given a Markov kernel space $(\mathbf{K}, E, F)$ along with a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ with $n \geq 2$ and a conditioning variable Z, $(\mathbf{K}, E, F)$ can be extended with a set of variables $\mathsf{Y}(D) := (\mathsf{Y}(i))_{i\in D}$ such that $\{\mathsf{Y}_i, \mathsf{Y}(D), \mathsf{D}_i\}$ is a parallel potential outcome submodel if and only if $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ is extendably exchangeable and deterministically reproducible conditional on Z.*

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

Let $e = (e_i)_{i\in[|D|]}$. Introduce the variable $\mathsf{Y}(i)$ for $i \in D$ such that $\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_{[D]}\mathsf{Z}]_{ez} = \mathbf{K}[\mathsf{Y}_D|\mathsf{D}_D\mathsf{Z}]_{ez}$ and introduce X_i, $i \in D$ such that $\mathbf{K}[\mathsf{X}_i|\mathsf{D}_i\mathsf{Z}\mathsf{Y}(D)]^{x_i}_{e_i z j_1 \dots j_{|D|}} =$

$\delta[j_{e_i}]^{x_i}$. Clearly $\{\mathsf{X}_{[n]}, \mathsf{D}_{[n]}, \mathsf{Y}(D)\}$ is a parallel potential outcome submodel. We aim to show that $\mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}] = \mathbf{K}[\mathsf{X}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$.

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

$$\mathbf{K}[\mathsf{X}_n|\mathsf{D}_n\mathsf{Z}]^x_{dz} = \sum_{y\in Y^{|D|}} \mathbf{K}[\mathsf{X}_{[n]}|\mathsf{D}_n\mathsf{Z}\mathsf{Y}(D)]^x_{dzy}\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_{[n]}\mathsf{Z}]^y_{dz} \tag{89}$$

$$= \sum_{y\in Y^{|D|}} \prod_{i\in[n]} \delta[y_{d_i}]^{x_i}\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_n\mathsf{Z}]^y_{dz} \tag{90}$$

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 $Q_d^C = [n] \setminus Q_d$, 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 $R_d^C = D \setminus R_d$.

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

$$\mathbf{K}[\mathsf{X}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]^x_{dz} = \sum_{y\in Y^{|D|}} \prod_{i\in Q_d} \delta[y_{d_i}]^{x_i}\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_{[n]}\mathsf{Z}]^y_{dz} \tag{91}$$

$$= \sum_{y'\in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}(R_d)\mathsf{Y}(R_d^C)|\mathsf{D}_{Q_d}\mathsf{D}_{Q_d^C}\mathsf{Z}]^{x_{Q_d}y'}_{d_{Q_d}d^C_{Q_d}z} \tag{92}$$

$$= \sum_{y'\in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}_{R_d}\mathsf{Y}_{R_d^C}|\mathsf{D}_{Q_d}\mathsf{D}_{Q_d^C}\mathsf{Z}]^{x_{Q_d}y'}_{dz} \tag{93}$$

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

$$\tag{94}$$

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. □

6 Appendix:see-do model representation

Modularise the treatment of probability

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

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

$$\mathbb{P}[\mathsf{X}\mathsf{D}\mathsf{Y}]^{ijk} = \mathbb{P}[\mathsf{X}]^i \mathbb{P}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{P}[\mathsf{Y}|\mathsf{X}\mathsf{D}]^k_{ij} \tag{95}$$

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

$$\mathbf{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{X}\mathsf{D}]^l_{kj} \tag{96}$$

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

$$\begin{aligned}
\mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj} &= \mathbf{T}\underline{\mathsf{X}'}^k_{hj} && (97)\\
&= \sum_l \mathbf{T}^{kl}_{hj} && (98)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^k && (99)\\
&= \mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj'} && (100)
\end{aligned}$$

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 95 to $\mathbb{P}_{h,\mathbf{S}}$:

$$\begin{aligned}
\mathbb{P}_{h,\mathbf{S}}[\mathsf{XDY}]^{ijk} &= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} && (101)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} && (102)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]^j_i \mathbf{T}[\mathsf{X}'\mathsf{Y}'|\mathsf{H}'\mathsf{D}']^{ik}_{hj} && (103)\\
&= \mathbf{S}^j_i \mathbf{T}[\mathsf{X}'\mathsf{Y}'|\mathsf{H}'\mathsf{D}']^{ik}_{hj} && (104)\\
&= \mathbf{S}^j_i \mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^i_{hj} \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} && (105)\\
&= \mathbf{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbf{S}^j_i \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} && (106)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i &= \mathbb{P}_{h,\mathbf{Q}}[\mathsf{X}]^i && (107)\\
\mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} &= \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \text{ for some version of } \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}] && (108)
\end{aligned}$$

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

$$\mathbf{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbf{Q}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}]^l_{kj} \tag{109}$$

Then by substitution of $\mathbf{Q}$ for $\mathbf{S}$ in Equation 101 and working through the same steps

$$\mathbb{P}_{h,\mathbf{S}}[\mathsf{XDY}]^{ijk} = \mathbf{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbf{Q}^j_i \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} \tag{110}$$

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

7 Appendix: Connection is associative

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

Define

$$\mathsf{I}_{K\cdot\cdot} := \mathsf{I}_K \setminus \mathsf{I}_L \setminus \mathsf{I}_J \tag{111}$$
$$\mathsf{I}_{KL\cdot} := \mathsf{I}_K \cap \mathsf{I}_L \setminus \mathsf{I}_J \tag{112}$$
$$\mathsf{I}_{K\cdot J} := \mathsf{I}_K \cap \mathsf{I}_J \setminus \mathsf{I}_L \tag{113}$$
$$\mathsf{I}_{KLJ} := \mathsf{I}_K \cap \mathsf{I}_L \cap \mathsf{I}_J \tag{114}$$
$$\mathsf{I}_{\cdot L\cdot} := \mathsf{I}_L \setminus \mathsf{I}_K \setminus \mathsf{I}_J \tag{115}$$
$$\mathsf{I}_{\cdot LJ} := \mathsf{I}_L \cap \mathsf{I}_J \setminus \mathsf{I}_K \tag{116}$$
$$\mathsf{I}_{\cdot\cdot J} := \mathsf{I}_J \setminus \mathsf{I}_K \setminus \mathsf{I}_L \tag{117}$$
$$\mathsf{O}_{K\cdot\cdot} := \mathsf{O}_K \setminus \mathsf{I}_N \setminus \mathsf{I}_J \tag{118}$$
$$\mathsf{O}_{KL\cdot} := \mathsf{O}_K \cap \mathsf{I}_L \setminus \mathsf{I}_J \tag{119}$$
$$\mathsf{O}_{K\cdot J} := \mathsf{O}_K \cap \mathsf{I}_J \setminus \mathsf{I}_L \tag{120}$$
$$\mathsf{O}_{KLJ} := \mathsf{O}_K \cap \mathsf{I}_L \cap \mathsf{I}_J \tag{121}$$
$$\mathsf{O}_{L\cdot} := \mathsf{O}_L \setminus \mathsf{I}_J \tag{122}$$
$$\mathsf{O}_{LJ} := \mathsf{O}_L \cap \mathsf{I}_J \tag{123}$$

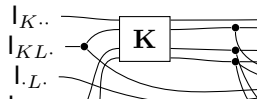
Also define

$$(\mathbf{P}, \mathsf{I}_P, \mathsf{O}_P) := \mathbf{K} \rightrightarrows \mathbf{L} \tag{124}$$
$$(\mathbf{Q}, \mathsf{I}_Q, \mathsf{O}_Q) := \mathbf{L} \rightrightarrows \mathbf{J} \tag{125}$$

Then

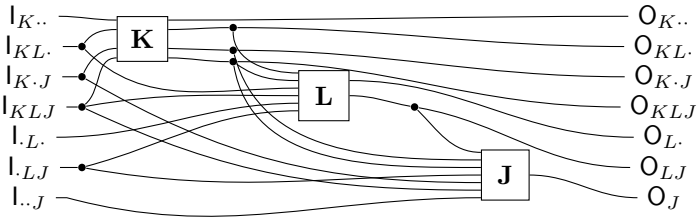
$$(\mathbf{K} \rightrightarrows \mathbf{L}) \rightrightarrows \mathbf{J} = \mathbf{P} \rightrightarrows \mathbf{J} \tag{126}$$

$$= \tag{127}$$

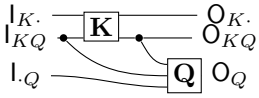


$$= \tag{128}$$

$$\overset{perm}{=} \tag{129}$$



$$= \tag{130}$$



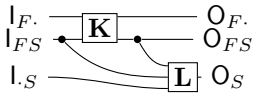
$$= \mathbf{K} \rightrightarrows (\mathbf{L} \rightrightarrows \mathbf{J}) \tag{131}$$

8 Appendix: String Diagram Examples

Recall the definition of *connection*:

Definition 8.1 (Connection)**.**

$$\mathbf{K} \rightrightarrows \mathbf{L} := \tag{132}$$



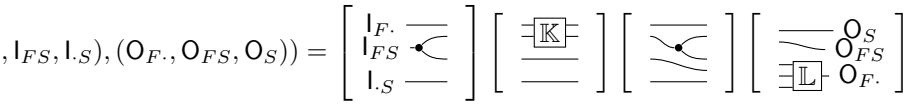
$$:= \mathbf{J} \tag{133}$$

$$\mathbf{J}^{zxw}_{yqr} = \mathbf{K}^{zx}_{yq}\mathbf{L}^{w}_{xqr} \tag{134}$$

Equation 132 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:

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



(135)

(136)

9 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 $(Y, \mathcal{Y})$

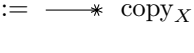
Note that this is a more general definition than the one used in the main paper; the version in the main paper is the restriction of this definition to finite sets.

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

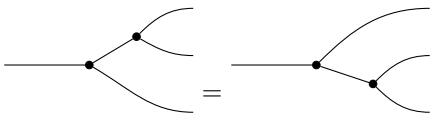
$$\delta(x, A) = \begin{cases} 1 & x \in A \\ 0 & \text{otherwise} \end{cases} \tag{137}$$

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

Definition 9.1. A Markov category C is a symmetric monoidal category in which every object $X \in C$ is equipped with a commutative comonoid structure given by a comultiplication $\mathrm{copy}_X : X \to X \otimes X$ and a counit $\mathrm{del}_X : X \to I$, depicted in string diagrams as

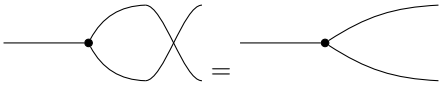
$\mathrm{del}_X :=$  $\mathrm{copy}_X \quad :=$ (138)

and satisfying the commutative comonoid equations



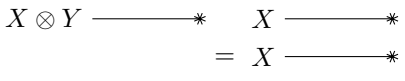
(139)

(140)

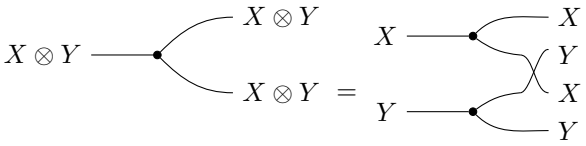


(141)

as well as compatibility with the monoidal structure

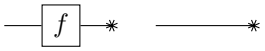


(142)



(143)

and the naturality of *del*, which means that



(144)

for every morphism f.

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

$$\mathbf{KL} : \mathsf{X} \to \Delta(\mathsf{Z}) \tag{145}$$

$$\text{defined by } (\mathbf{KL})(x, A) = \int_Y \mathbf{L}(y, A)\mathbf{K}(x, dy) \tag{146}$$

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

$$\mathbf{K} \otimes \mathbf{L} : (\mathsf{X}, \mathsf{W}) \to \Delta(\mathsf{Y}, \mathsf{Z}) \tag{147}$$

$$\text{defined by } \mathbf{K} \otimes \mathbf{L}(x, w, A \times B) = \mathbf{K}(x, A)\mathbf{L}(w, B) \tag{148}$$

The identity map is

$$\mathrm{Id}_{\mathsf{X}} : \mathsf{X} \to \Delta(\mathsf{X}) \tag{149}$$

$$\text{defined by} (\mathrm{Id}_X)(x, A) = \delta(x, A) \tag{150}$$

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

$$\text{swap}_{\mathsf{X},\mathsf{Y}} : (\mathsf{X},\mathsf{Y}) \to \Delta(\mathsf{Y},\mathsf{X}) \tag{151}$$

$$\text{defined by} (\text{swap}_{\mathsf{X},\mathsf{Y}})(x, y, A \times B) = \delta(x, B)\delta(y, A) \tag{152}$$

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

$$\text{copy}_{\mathsf{X}} : \mathsf{X} \to \Delta(\mathsf{X},\mathsf{X}) \tag{153}$$

$$\text{defined by} (\text{copy}_X)(x, A \times B) = \delta_x(A)\delta_x(B) \tag{154}$$

and the erase map by

$$\text{del}_{\mathsf{X}} : \mathsf{X} \to \Delta(*) \tag{155}$$

$$\text{defined by} (\text{del}_X)(x, A) = \delta(*, A) \tag{156}$$

$$\tag{157}$$

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 9.2 (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 139 to 144 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. P. Dawid. Causal Inference without Counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424, June 2000. ISSN 0162-1459. doi: 10.1080/01621459.2000.10474210. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.2000.10474210. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.2000.10474210.

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.

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.

James M. Robins and Sander Greenland. Causal Inference Without Counterfactuals: Comment. *Journal of the American Statistical Association*, 95 (450):431–435, 2000. ISSN 0162-1459. doi: 10.2307/2669381. URL http://www.jstor.org/stable/2669381. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

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: