# Understanding Causal Primitives Using Modular Probability

David Johnston

October 8, 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.

Vague variables are crucial for expressing the fact that a given model says something about the real world. If I'm interested in the heights of people in Chicago, then I can't simply construct a set and define some formal variables – I must consider some family of things that actually have some correspondence with people's heights in Chicago.

One way to view vague variables is as an existential claim: there is *some* sample space and *some* function that behaves exactly like the thing I have vaguely indicated in the definition. If you and I start measuring people's heights and I tell you "$\mathsf{X}_i$ represents the $i$th person's height", then we can go and measure someone's height and in the end we both somehow decide decide that the appropriate value of $\mathsf{X}_1$ is, say, 1.78 metres. Whatever actually happens to get from "let's measure people's height" to $\mathsf{X}_1 = 1.78$ is extremely hard to explain, but if we can be confident of always agreeing on the end result then the definition appears to have been a success. Furthermore, given that we both come to the same conclusion based on observing the same experiment perhaps this observation process, however it actually works, can be modeled with a some function on some sample space.

There are also other kinds of things we also use variables to model. For example, "latent variables" are by definition unobserved, so we cannot posit a story about them being defined by some "process of observation". Latent variables are often introduced in order to model relationships between observed variables, for example "if the average height is $\mathsf{M}$ and the variance $\mathsf{V}$, then $\mathsf{X}_1$ will be distributed according to $\mathcal{N}(\mathsf{M}, \mathsf{V})$". We can extend the existential claim: there is some sample space and some collection of variables on it that describe the process of observing and supposing.

If all we need is *some* sample space does the job we need, we are free to choose the most convenient one. Thus if all we care about are variables $\mathsf{X}$ with codomain $X$ and $\mathsf{Y}$ with codomain $Y$, and they can mutually take any pair of values, then $X \times Y$ might be sufficient.

However, for some questions we do need to be mindful of the complexities of vague variables, particularly when we are doing causal inference. Firstly, variables representing *choices* may have a different scope to most types of variables. Consider two vague variables: $\mathsf{X}$ represents "whether I get regular exercise next month" and $\mathsf{D}$ represents "whether I decide to get regular exercise next month". Many people might be able to observe my exercise next month, and therefore might be expected to agree on the value of $\mathsf{X}$ when all is said and done. On the other hand, most people won't be able to say what value $\mathsf{D}$ should take, as whether or not I contemplated a choice and what choice I took is information that will usually be available only to me. However, I will be able to say what value $\mathsf{D}$ took, and I will agree with my previous judgement if I ask myself this on multiple occasions.

Secondly, the fact that a practical sample space might be constructible for a given set of vague variables does not imply that there is a practical way to consider the set of *all* variables. Something like "for every variable $\mathsf{Z}$, $\mathsf{Z}$ is not a common cause of $\mathsf{X}$ and $\mathsf{Y}$" would seem to require consideration of all variables on the "actual sample space", whatever exactly that is, rather than a convenient summary.

## 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 2.15).

We will be following the general approach discussed above, assuming that there is some measurable sample space $\Omega$ and that variables are measurable functions defined on $\Omega$. It is also often standard to assume that we have a *probability space* $(\mathbb{P}, (\Omega, \mathcal{F}))$, where $\mathbb{P}$ is a $\sigma$-additive measure on $(\Omega, \mathcal{F})$ with $\mathbb{P}(\Omega) = 1$. For causal models, we are often interested in collections of probability measures $\{\mathbb{P}_\alpha | \alpha \in A\}$ or in models that feature what Constantinou and Dawid (2017) refers to as "non-stochastic variables".

More generally, we are also interested in non-standard compositions of functions representing conditional probabilities. To illustrate the motivation for this, consider "truncated factorisation", a linchpin operation in causal graphical models. Suppose we have a causal Bayesian network $(\mathbb{P}^{\mathsf{XYZ}}, \mathcal{G})$ where $\mathcal{G}$ is a Directed Acyclic Graph that features the edges $\mathsf{X} \longrightarrow \mathsf{Y}$ and $\mathsf{X} \longleftarrow \mathsf{Z} \longrightarrow \mathsf{Y}$. Then the result of "setting $\mathsf{X}$ to $x$" is represented by a new probability measure $\mathbb{P}_x$ that is required to obey the truncated factorisation (Pearl, 2009, page 24):

$$\mathbb{P}_x^{\mathsf{XYZ}}(x', y, z) = \mathbb{P}^{\mathsf{Y|XZ}}(y|x, z)\mathbb{P}^{\mathsf{Z}}(z)[\![x = x']\!] \tag{1}$$

There is no special siginificance in standard probability theory to the expression on the right hand side of Equation 1. This equation is an example of an operation that combines an irregular pair of probability models and returns a probability model *defined on the same variables*. Here we consider in general when it is possible to combine irregular collections of conditional or marginal probabilities (we refer to these as *submodels*) and get a probability model *over the same variables*.

The main challenge with combining submodels is that, as we will show, this operation is "unsafe" in the sense that the combination of two well-formed submodels might return a submodel that is not itself well-formed. For a simple example of this, suppose in Equation 1 that $\mathsf{X}$ represents a person's body mass index while $\mathsf{Z} = (\mathsf{W}, \mathsf{H})$ represents the pair of their weight and height. The collection $\{\mathbb{P}_x^{\mathsf{XYZ}} | x \in X\}$ then models a situation where we change a patient's body mass index while leaving their weight and height unchanged, which is not a particularly useful model, and does not abide by any reasonable definition of "a variable representing body mass index".

We discuss some conditions under which the result of composition is 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 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 1 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 $\sigma$-algebra, we can represent a Markov kernel $\mathbf{K}$ as a $|X| \times |Y|$ matrix where $\sum_{y \in Y} \mathbf{K}_x^y = 1$ for every $x \in X$. We will give Markov kernels the signature $\mathbf{K} : X \rightarrowtriangle Y$ to indicate that they map from $X$ to probability distributions on $Y$.

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

$$\mathbf{K} := -\boxed{\mathbf{K}}- \tag{2}$$

$$\mathbf{P} := \lhd\!\mathbf{P}- \tag{3}$$

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

$$\mathbf{LM} := -\boxed{\mathbf{K}}-\boxed{\mathbf{M}}- \tag{4}$$

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

The tensor product is represeted by drawing kernels in parallel:

$$\mathbf{K} \otimes \mathbf{L} := \begin{array}{c} W \boxed{\mathbf{K}}\, Y \\ X\, \boxed{\mathbf{L}}\, Z \end{array} \tag{5}$$

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

A description of these kernels follows.

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

$$\mathrm{id}_X := \; X \text{-} X \tag{6}$$

We choose a particular 1-element set $\{*\}$ that acts as the identity in the sense that $\{*\} \times A = A \times \{*\} = A$ for any set $A$. The erase map $\mathrm{del}_X : X \rightarrowtriangle \{*\}$ defined by $(\mathrm{del}_X)_x^* = 1$ is a Markov kernel that "discards the input" (we will later use it for marginalising joint distributions). It is drawn as a fuse:

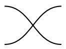
$$\mathrm{del}_X := \quad -\!\!\ast \; X \tag{7}$$

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

$$\mathrm{copy}_X := \; X -\!\!\!\bullet\!\!\!< \begin{matrix} X \\ X \end{matrix} \tag{8}$$

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

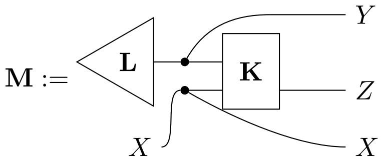
$$\mathrm{swap}_X := \tag{9}$$



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

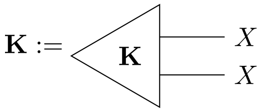
The Markov kernels introduced in the previous section can be though of as "conditional probability distributions without variables". We can use these to represent an operation very similar to Equation 1. Note that $P^{\mathsf{Y}|\mathsf{XZ}}$ must be represented by a Markov kernel $\mathbf{K} : X \times Z \rightarrowtriangle Y$ and $\mathbb{P}^{\mathsf{Z}}$ by a Markov kernel $\mathbf{L} \in \Delta(Z)$. Then we can define a Markov kernel $\mathbf{M} : X \rightarrowtriangle X \times Z$ representing $x \mapsto \mathbb{P}_x^{\mathsf{YZ}}(y, z)$ by



(10)

There is, however, a key difference between Equation 10 and Equation 1: the Markov kernels in the latter equation describe the distribution of particular variables, while the former equation describes Markov kernels only.

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



(11)

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

## 2.5 Composition and probability with variables

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

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

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

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

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

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

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

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

$$\mathbf{K} := \begin{array}{c}\mathsf{A}_1\\ \mathsf{A}_2\end{array} \boxed{\mathbf{K}} \begin{array}{c}\mathsf{B}_1\\ \mathsf{B}_2\end{array} \tag{12}$$

or

$$\mathbf{K} = (\mathsf{A}_1, \mathsf{A}_2) \boxed{\mathbf{K}[\mathbf{L}]} (\mathsf{B}_1, \mathsf{B}_2) \tag{13}$$

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

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

We define $\mathrm{Id}_{\mathsf{X}}$ 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 $\mathsf{X}$ takes values in $\{0, 1\}$ we can propose an indexed Markov kernel $\mathbf{K} : \mathsf{X} \rightarrowtriangle \mathsf{X}$ with $\mathbf{K}'^{b}_{a} = 0.5$ for all $a, b$. However, this is not a useful model of the variable $\mathsf{X}$ – it expresses something like "if we know the value of $\mathsf{X}$, then we belive that $\mathsf{X}$ 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".

**Definition 2.3** (Model)**.** An indexed Markov kernel $(\mathbf{K}', \mathsf{X}, \mathsf{Y})$ is a *model* if it is *consistent*, which means it assigns probability 0 to contradictions:

$$f_{\mathsf{X}}^{-1}(a) \cap f_{\mathsf{Y}}^{-1}(b) = \emptyset \implies \left(\mathbf{K}'^{b}_{a} = 0\right) \vee \left(f_{\mathsf{X}}^{-1}(a) = \emptyset\right) \tag{14}$$

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

Here a contradiction is a simultaneous assignment of values to the variables $\mathsf{X}$ and $\mathsf{Y}$ such that there is no value of $\omega$ 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.

Consistency has some interesting implications. For a start, given a probability model of the sample space, there is a unique corresponding probability model of $\mathsf{X}$ is given by the pushforward of $\mathsf{X}$ (corollary 2.6). Thus it motivates the definition of the pushforward measure as the distribution of $\mathsf{X}$ induced by the sample space model.

Consistency also implies that for any $\mathbf{K} : \mathsf{X} \rightarrowtriangle \mathsf{Y}$, if $f_{\mathsf{Y}} = g \circ f_{\mathsf{X}}$ then $\mathbf{K}_x^{g(x)} = 1$. A particularly simple case of this is a model $\mathbf{L} : \mathsf{X} \rightarrowtriangle \mathsf{X}$, which must be such that $\mathbf{L}_x^x = 1$. 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.

The consistency condition is also needed to ensure sample space models with nontrival domains such as $\mathbf{K} : \mathsf{X} \rightarrowtriangle \mathrm{Id}_{\mathit{\Omega}}$ conform to expectations imposed by the random variables. We might want to consider such models if we want to include "nonstochastic variables" (Constantinou and Dawid, 2017).

**Lemma 2.4** (Uniqueness of models with the sample space as a domain)**.** *For any $\mathsf{X} : \mathit{\Omega} \to A$, there is a unique model $\mathbf{X} : Id_{\mathit{\Omega}} \rightarrowtriangle \mathsf{X}$ given by $\mathbf{X} := (\mathsf{X}, Id_{\mathit{\Omega}}, \mathsf{X})$.*

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

For any $\omega \in \mathit{\Omega}$, $a \in A$

$$\begin{aligned}
\max_{\omega\in\mathit{\Omega}} (\mathrm{Id}_{\mathit{\Omega}}, \mathsf{X})_{\omega}^{\omega',a} &= \max_{\omega\in\mathit{\Omega}} \llbracket \omega = \omega' \rrbracket \llbracket \omega = f_{\mathsf{X}}(a) \rrbracket && (15)\\
&= \llbracket \omega = f_{\mathsf{X}}(a) \rrbracket && (16)\\
&= \mathbf{X}_{\omega}^{a} && (17)
\end{aligned}$$

Thus $\mathbf{X}$ satisfies consistency.

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

$$\begin{aligned}
\max_{\omega\in\mathit{\Omega}} (\mathrm{Id}_{\mathit{\Omega}}, \mathsf{X})_{\omega}^{\omega',a} &= \max_{\omega\in\mathit{\Omega}} \llbracket \omega = \omega' \rrbracket \llbracket \omega = f_{\mathsf{X}}(b) \rrbracket && (18)\\
&= \llbracket \omega = f_{\mathsf{X}}(b) \rrbracket && (19)\\
&= 0 && (20)\\
&< \mathbf{K}_{\omega}^{b} && (21)
\end{aligned}$$

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

**Lemma 2.5** (Pushforward models)**.** *Given any model $\mathbf{P} : \mathsf{Y} \rightarrowtriangle Id_{\mathit{\Omega}}$, there is a unique model $\mathbf{P}^{\mathsf{X}|\mathsf{Y}} : \mathsf{Y} \rightarrowtriangle \mathsf{X}$ such that $\mathbf{P}^{\mathsf{X}|\mathsf{Y}} = \mathbf{P}\mathbf{Q}$ for some $\mathbf{Q} : Id_{\mathit{\Omega}} \to \mathsf{X}$, and it is given by $(\mathbf{P}^{\mathsf{X}|\mathsf{Y}})_b^a = \sum_{\omega\in f^{-1}(a)} \mathbf{P}_b^{\omega}$.*

*Proof.* As $\mathbf{X} := (\mathsf{X}, \mathrm{Id}_\Omega, \mathsf{X})$ is the unique model $\mathrm{Id}_\Omega \to \mathsf{X}$, it must be the case that $\mathbf{P}^{\mathsf{X}|\mathsf{Y}} = \mathbf{PX}$. Suppose $\mathsf{X} : \Omega \rightarrowtriangle A$ and $\mathsf{Y} : \Omega \rightarrowtriangle B$. Then for any $a \in A$, $b \in B$

$$(\mathbf{PX})^a_b = \sum_{\omega\in\Omega} \mathbf{P}^\omega_b \mathsf{X}^a_\omega \tag{22}$$

$$= \sum_{\omega\in\Omega} \mathbf{P}^\omega_b \llbracket a = f_{\mathsf{X}}(\omega)\rrbracket \tag{23}$$

$$= \sum_{\omega\in f^{-1}(a)} \mathbf{P}^\omega_b \tag{24}$$

□

**Corollary 2.6** (Pushforward probability model)**.** *Given any probability model* $\mathbf{P} : \mathsf{I} \rightarrowtriangle Id_\Omega$*, there is a unique model* $\mathbf{P}^{\mathsf{X}} : \mathsf{I} \rightarrowtriangle \mathsf{X}$ *such that* $\mathbf{P}^{\mathsf{X}} = \mathbf{PQ}$ *for some* $\mathbf{Q} : Id_\Omega \to \mathsf{X}$*, and it is given by* $(\mathbf{P}^{\mathsf{X}})^a_b = \sum_{\omega\in f^{-1}(a)} \mathbf{P}^\omega_b$.

*Proof.* Apply Lemma 2.5 to a model $\mathbf{P} : \mathsf{I} \rightarrowtriangle \mathrm{Id}_\Omega$. □

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

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

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

$$\mathbf{K}'^{y,y',z}_x = \llbracket y = y' \rrbracket \mathbf{L}'^{y,z}_x \tag{25}$$

$$\tag{26}$$

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

$$(\mathsf{X}, \mathsf{Y}, \mathsf{Y}, \mathsf{Z})^{x,y,y',z}_\omega = \llbracket f_{\mathsf{Y}}(\omega) = y\rrbracket \llbracket f_{\mathsf{Y}}(\omega) = y'\rrbracket (\mathsf{X}, \mathsf{Z})^{x,z}_\omega \tag{27}$$

$$= \llbracket y = y' \rrbracket \llbracket f_{\mathsf{Y}}(\omega) = y\rrbracket (\mathsf{X}, \mathsf{Z})^{x,z}_\omega \tag{28}$$

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 $\mathbf{L}$ is a model follows from the assumption that $\mathbf{K}$ is. Then

$$\mathbf{K}'^{y,y',z}_x = \llbracket y = y' \rrbracket \mathbf{L}'^{y,z}_x \tag{29}$$

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

$$\llbracket y = y' \rrbracket \mathbf{L}'^{y,z}_x > 0 \implies y = y' \wedge \mathbf{L}'^{y,z}_x > 0 \tag{30}$$

$$\implies \left(f^{-1}_{\mathsf{X}}(x) = \emptyset\right) \vee \left(f^{-1}_{\mathsf{X}}(x) \cap f^{-1}_{\mathsf{Y}}(y) \cap f^{-1}_{\mathsf{Y}}(y) \cap f^{-1}_{\mathsf{Z}}(z) \neq \emptyset\right) \tag{31}$$

$$\tag{32}$$

□

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

**Lemma 2.8** (Copies shared between input and output are identical)**.** *For any* $\mathbf{K} : (\mathsf{X}, \mathsf{Y}) \rightarrowtriangle (\mathsf{X}, \mathsf{Z})$*,* $\mathbf{K}$ *is a model iff there exists some* $\mathbf{L} : (\mathsf{X}, \mathsf{Y}) \rightarrowtriangle \mathsf{Z}$ *such that*

$$\mathbf{K}'^{x',z}_{x,y} = \llbracket x = x' \rrbracket \mathbf{L}^{z}_{'x,y} \tag{33}$$

*Proof.* $\Longrightarrow$ For any $\omega, x, y, y', z$:

$$\begin{aligned}(\mathsf{X}, \mathsf{Y}, \mathsf{Y}, \mathsf{Z})^{x,y,y',z}_{\omega} &= \llbracket f_{\mathsf{Y}}(\omega) = y \rrbracket \llbracket f_{\mathsf{Y}}(\omega) = y' \rrbracket (\mathsf{X}, \mathsf{Z})^{x,z}_{\omega} &(34)\\ &= \llbracket y = y' \rrbracket \llbracket f_{\mathsf{Y}}(\omega) = y \rrbracket (\mathsf{X}, \mathsf{Z})^{x,z}_{\omega} &(35)\end{aligned}$$

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

$$\mathbf{K}'^{x',z}_{x,y} = \llbracket x = x' \rrbracket \mathbf{L}'^{z}_{x,y} \tag{36}$$

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

$$\begin{aligned}\llbracket x = x' \rrbracket \mathbf{L}'^{z}_{x,y} > 0 &\implies x = x' \wedge \mathbf{L}'^{z}_{x,y} > 0 &(37)\\ &\implies \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) = \emptyset\right) \vee \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) \cap f_{\mathsf{Z}}^{-1}(z) \neq \emptyset\right) &(38)\\ & &(39)\end{aligned}$$

□

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

**Definition 2.9** (Variation independence)**.** Two variables $\mathsf{X} : \mathit{\Omega} \rightarrowtriangle X$ and $\mathsf{Y} : \mathit{\Omega} \rightarrowtriangle Y$ are variation independent, written $\mathsf{X} \perp_v \mathsf{Y}$, if for all $y \in f_{\mathsf{Y}}(\mathit{\Omega}) R(f_{\mathsf{Y}})$

$$f_{\mathsf{Y}}(\mathit{\Omega}) \times f_{\mathsf{X}}(\mathit{\Omega}) = \{(f_{\mathsf{Y}}(\omega), f_{\mathsf{X}}(\omega)) | \omega \in \mathit{\Omega}\} \tag{40}$$

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

**Lemma 2.10** (Consistency via variation conditional independence)**.** *Given an indexed Markov kernel* $\mathbf{K} : \mathsf{X} \rightarrowtriangle \mathsf{Y}$ *with* $\mathsf{X} : \mathit{\Omega} \rightarrowtriangle X$ *and* $\mathsf{Y} : \mathit{\Omega} \rightarrowtriangle Y$*, if* $f_{\mathsf{Y}}$ *is surjective and* $\mathsf{Y} \perp_v \mathsf{X}$ *then* $\mathbf{K}$ *is a model.*

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

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

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

**Lemma 2.11** (Consistency via positive models)**.** *Given a model* $\mathbf{K} : \mathsf{X} \rightarrow (\mathsf{Y}, \mathsf{Z})$*, if an indexed Markov kernel* $\mathbf{L} : (\mathsf{X}, \mathsf{Y}) \rightarrow \mathsf{Z}$ *has the property* $\mathbf{K}'^{yz}_x = 0 \implies \mathbf{L}'^z_{xy} = 0$ *then* $\mathbf{L}$ *is also a model.*

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

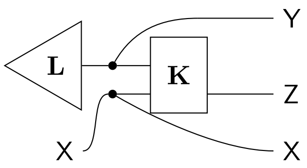
$$\begin{aligned}\mathbf{L}'^z_{xy} > 0 &\implies \mathbf{K}'^{yz}_x > 0 &(41)\\ &\implies \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) \cap f_{\mathsf{Z}}^{-1}(z) \neq \emptyset\right) \vee \left(f_{\mathsf{X}}^{-1}(x) = \emptyset\right) &(42)\\ &\implies \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) \cap f_{\mathsf{Z}}^{-1}(z) \neq \emptyset\right) \vee \left(f_{\mathsf{X}}^{-1}(x) \cap f_{\mathsf{Y}}^{-1}(y) = \emptyset\right) &(43)\end{aligned}$$

□

## 2.6 Truncated factorisation with variables

At this point, we can represent Equation 1 using models. Suppose $P^{\mathsf{Y}|\mathsf{XZ}}$ is an model $\mathbf{K} : (\mathsf{X}, \mathsf{Z}) \rightarrow \mathsf{Y}$ and $\mathbb{P}^{\mathsf{Z}}$ an model $\mathbf{L} : \{*\} \rightarrow \mathsf{Z}$. Then we can define an indexed Markov kernel $\mathbf{M} : \mathsf{X} \rightarrow \mathsf{X}, \mathsf{Z}$ representing $x \mapsto \mathbb{P}^{\mathsf{YZ}}_x(y, z)$ by

$\mathbf{M} :=$



(44)

Equation 44 is almost identical to Equation 10, except it now specifies which variables each measure applies to, not just which sets they take values in. Like the original Equation 1, there is no guarantee that $\mathbf{M}$ is actually a model. If $f_{\mathsf{X}} = g \circ f_{\mathsf{Z}}$ for some $g : Z \to X$ and $X$ has more than 1 element, then the rule of consistency will rule out the existence of any such model.

If we want to use $\mathbf{M}$, we want it at minimum to satisfy the consistency condition. One approach we could use is to check the result using Lemmas 2.7 to 2.11, although note that 2.10 and 2.11 are sufficient conditions, not necessary ones.

## 2.7 Sample space models and submodels

Instead of trying to assemble probability models as in Equation 44, we might try to build probability models in a manner closer to the standard setup – that is, we start with a sample space model (or a collection of sample space models) and work with marginal and conditional probabilities derived from these, without using any non-standard model assemblies.

A sample space model is any model $\mathbf{K} : \mathsf{X} \rightarrowtriangle \mathrm{Id}_{\mathit{\Omega}}$. 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 2.12** (Existence of sample space model)**.** *Given any model* $\mathbf{K} : \mathsf{X} \rightarrowtriangle \mathsf{Y}$*, there is a sample space model* $\mathbf{L} : \mathsf{X} \rightarrowtriangle \mathit{Id}_{\mathit{\Omega}}$ *such that, defining* $\mathbf{Y} := (\mathsf{Y}, \mathit{Id}_{\mathit{\Omega}}, \mathsf{Y})$*,* $\mathbf{LY} = \mathbf{K}$*.*

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

$$\mathbf{L}'^{\omega}_{a} = \begin{cases} 0 & \text{if } f_{\mathsf{Y}}^{-1}(b) \cap f_{\mathsf{X}}^{-1}(a) = \emptyset \\ \mathbf{K}'^{b}_{a}[\![\omega = \omega_b]\!] & \text{for some } \omega_b \in f_{\mathsf{Y}}^{-1}(b) \text{ if } f_{\mathsf{X}}^{-1}(a) = \emptyset \\ \mathbf{K}'^{b}_{a}[\![\omega = \omega_{ab}]\!] & \text{for some } \omega_{ab} \in f_{\mathsf{Y}}^{-1}(b) \cap f_{\mathsf{X}}^{-1}(a) \text{ otherwise} \end{cases} \tag{45}$$

Note that for all $a \in A$, $\sum_{\omega \in \mathit{\Omega}} \mathbf{L}'^{\omega}_{a} = \sum_{b \in B} \mathbf{K}'^{b}_{a} = 1$.

By construction, $(\mathbf{L}', \mathrm{Id}_{\mathit{\Omega}}, \mathsf{X})$ is free of contradiction. In addition

$$(\mathbf{L}'\mathsf{Y})^{b}_{a} = \sum_{\omega \in \mathit{\Omega}} \mathbf{L}'^{\omega}_{a} \mathsf{Y}^{b}_{\omega} \tag{46}$$

$$= \sum_{\omega \in f_{\mathsf{Y}}^{-1}(b)} \mathbf{L}'^{\omega}_{a} \tag{47}$$

$$= \begin{cases} 0 & f_{\mathsf{Y}}^{-1}(b) \cap f_{\mathsf{X}}^{-1}(a) = \emptyset \\ \mathbf{K}'^{b}_{a} & \text{otherwise} \end{cases} \tag{48}$$

$$\implies (\mathbf{L}'\mathsf{Y}) = \mathbf{K}' \tag{49}$$

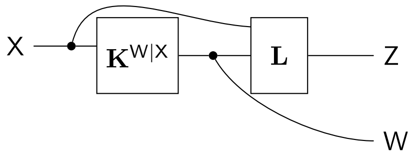
□

**Definition 2.13** (Pushforward model)**.** For any variables $\mathsf{X} : \mathit{\Omega} \rightarrowtriangle A$, $\mathsf{Y} : \mathit{\Omega} \rightarrowtriangle B$ and any sample space model $\mathbf{K} : \mathsf{X} \rightarrowtriangle \mathrm{Id}_{\mathit{\Omega}}$, the pushforward $\mathbf{K}^{\mathsf{Y}|\mathsf{X}} := \mathbf{KX}$ where $\mathbf{X} := (\mathsf{X}, \mathrm{Id}_{\mathit{\Omega}}, \mathsf{X})$.

The fact that the pushforward is a model is proved in Lemma 2.5. We employ the slightly more familiar notation $\mathbf{K}^{\mathsf{Y}|\mathsf{X}}(y|x) \equiv (\mathbf{K}'^{\mathsf{Y}|\mathsf{X}})^{y}_{x}$.

**Definition 2.14** (Submodel)**.** Given $\mathbf{K} : \mathsf{X} \rightarrowtriangle \mathrm{Id}_{\mathit{\Omega}}$ and $\mathbf{L} : \mathsf{W}, \mathsf{X} \rightarrowtriangle \mathsf{Z}$, $\mathbf{L}$ is a submodel of $\mathbf{K}$ if

$$\mathbf{K}^{\mathsf{Z},\mathsf{W}|\mathsf{Y}} = \tag{50}$$



$$(\mathbf{K}^{\mathsf{Z},\mathsf{W}|\mathsf{Y}})_x^{w,z} = (\mathbf{K}^{\mathsf{W}|\mathsf{Y}})_x^w \mathbf{L}_{w,x}^z \tag{51}$$

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

**Lemma 2.15** (Submodel existence)**.** *For any model* $\mathbf{K} : \mathsf{X} \rightarrowtriangle \mathrm{Id}_\Omega$ *(where $\Omega$ is a finite set),* $\mathsf{W}$ *and* $\mathsf{Y}$*, there exists a submodel* $\mathbf{L} : (\mathsf{W},\mathsf{X}) \rightarrowtriangle \mathsf{Y}$*.*

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

$$\mathbf{L}'^{y}_{wx} = \frac{\mathbf{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w,y|x)}{\mathbf{K}^{\mathsf{W}|\mathsf{X}}(w|x)} \qquad \forall x,w: \text{ the denominator is positive} \tag{52}$$

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 $\mathbf{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w,y|x) > 0 \implies \mathbf{L}'^{y}_{wx} > 0$. Furthermore, where $\mathbf{K}^{\mathsf{W}|\mathsf{X}}(w|x) = 0$, we either have $f_{\mathsf{W}}^{-1}(w) \cap f_{\mathsf{X}}^{-1}(x) = \emptyset$ or we can choose some $\omega_{wx} \in f_{\mathsf{W}}^{-1}(w) \cap f_{\mathsf{X}}^{-1}(x)$ and let $\mathbf{L}'^{f_{\mathsf{Y}}(\omega_{wx})}_{wx} = 1$. Thus $\mathbf{L}'$ can be chosen such that $\mathbf{L}$ is a model (but this is not automatic).

Then

$$\begin{aligned}
\mathbf{K}^{\mathsf{W}|\mathsf{X}}(w|x)\mathbf{L}'^{y}_{xw} &= \mathbf{K}^{\mathsf{W}|\mathsf{X}}(w|x)\frac{\mathbf{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w,y|x)}{\mathbf{K}^{\mathsf{W}|\mathsf{X}}(w|x)} && \text{if } \mathbf{K}^{\mathsf{W}|\mathsf{X}}(w|x) > 0 && (53)\\
&= \mathbf{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w,y|x) && \text{if } \mathbf{K}^{\mathsf{W}|\mathsf{X}}(w|x) > 0 && (54)\\
&= 0 && \text{otherwise} && (55)\\
&= \mathbf{K}^{\mathsf{W},\mathsf{Y}|\mathsf{X}}(w,y|x) && \text{otherwise} && (56)
\end{aligned}$$

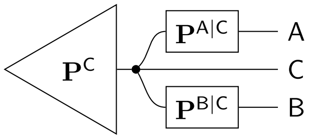
□

## 2.8 Conditional independence

We define conditional independence in the following manner:

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

$$\mathbf{P}^{\mathsf{ABC}} = \tag{57}$$

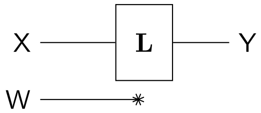

$\mathbf{P}^{\mathsf{C}}$
$\mathbf{P}^{\mathsf{A}|\mathsf{C}}$
A
C
$\mathbf{P}^{\mathsf{B}|\mathsf{C}}$
B

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

This definition is inapplliccable 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:

$\mathbf{K} =$



(58)

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 $\mathsf{X}$ taking values in $X$, and in response to this we can select any decision, modelled by $\mathsf{D}$ taking values in $D$. Picking a decision, or mixture of decisions, is a *strategy* $\alpha$, modelled by $\mathbf{S}_\alpha : \mathsf{X} \to \Delta(\mathsf{D})$ (In principle we can make $\alpha$ a variable too, but in that case we need to generalise our theory of probability models to cover continuously valued variables). We are interested in some defined collection of things that happen at some point after we have taken our decision; these will be modelled by the variable $\mathsf{Y}$ which we will call the consequences.

Clearly, for different observations and strategy choices we should expect different consequences.

We will represent our uncertain answers to the questions $\mathsf{X}$, $\mathsf{D}$ and $\mathsf{Y}$ with probability distributions. We will allow for multiple probability distributions to be entertained as an answer; let hypotheses $\mathsf{H}$ represent the question "which model best captures this problem?", taking values in $H$. Then for each strategy $\mathbf{S}_\alpha$, our forecast will be represented by a joint probability $\mathbf{P}_\alpha \equiv \mathbf{P}_\alpha^{\mathsf{XDY}|\mathsf{H};\mathbf{P}_\alpha} : \mathsf{H} \to \Delta(\mathsf{X}, \mathsf{D}, \mathsf{Y})$. Because observations come before we execute our strategy, we might assume that they will be unchanged by any choice of strategy: $\mathbf{P}_\alpha^{\mathsf{X}|\mathsf{H}} = P_\beta^{\mathsf{X}|\mathsf{H}}$ for all $\alpha, \beta$. We expect to choose a decision precisely in line with the strategy under consideration: $\mathbf{P}_\alpha^{\mathsf{D}|\mathsf{X}} = \mathbf{S}_\alpha$. Finally, our answer to $\mathsf{Y}$ will be the same under any strategy supposing we have the same observation, decision and hypothesis: $\mathbf{P}_\alpha^{\mathsf{Y}|\mathsf{HD}} = P_\beta^{\mathsf{Y}|\mathsf{HD}}$ for all $\alpha, \beta$.

Under these assumptions, there exists $\mathbf{T}[\mathsf{XY}|\mathsf{HD}] \in \mathcal{M}$ with $\mathsf{X} \perp\!\!\!\perp_{\mathbf{T}} \mathsf{D}|\mathsf{H}$ such that for all $\alpha$,

$$\mathbf{P}_\alpha[\mathsf{XDY}|\mathsf{H}] \overset{krn}{=} \mathbf{T}[\mathsf{X}|\mathsf{H}] \rightrightarrows \mathbf{S}_\alpha[\mathsf{D}|\mathsf{X}] \rightrightarrows \mathbf{T}[\mathsf{Y}|\mathsf{XHD}] \tag{59}$$

The proof is given in Appendix 6. Note that $\mathbf{T}[\mathsf{X}|\mathsf{H}]$ exists by virtue of the fact $\mathsf{X} \perp\!\!\!\perp_{\mathbf{T}} \mathsf{D}|\mathsf{H}$. While this independence is what enables Equation 59, in general $\mathsf{X} \not\perp\!\!\!\perp_{\mathbf{P}_\alpha} \mathsf{D}|\mathsf{H}$, so $\mathbf{T}$ cannot be a disintegration of $\mathbf{P}_\alpha$. Modular probability allows us to specify $\mathbf{T}$, which we call a *see-do model*, as a partial forecast to be completed with a strategy $\mathbf{S}_\alpha$ while also being able to use consistent names for variables that represent the same things (observations, decisions, consequences, hypotheses) whether their distributions are given by $\mathbf{P}_\alpha$, $\mathbf{T}$, which are mutually incompatible conditional probabilities.

## 3.1 See-do models and classical statistics

A *statistical model* (or *statistical experiment*) is a collection of probability distributions indexed by some set $\Theta$. We can observe that $\{\mathbf{T}[\mathsf{X}|\mathsf{H}]_h\}_{h\in H}$ is a collection of probability distributions indexed by $H$.

In statistical decision theory, as introduced by Wald (1950), we are given a statistical experiment $\{\mathbb{P}_\theta \in \Delta(X)\}_\Theta$, a decision set $D$ and a loss $l : \Theta \times D \to \mathbb{R}$. A strategy $\mathbf{S}_\alpha : X \to \Delta(D)$ is evaluated according to the risk functional $R(\theta, \mathbf{S}_\alpha) = \sum_{x\in X} \sum_{d\in D} \mathbb{P}_\theta^x (S_\alpha)_x^d l(h, d)$.

Suppose we have a see-do model $\mathbf{T}[\mathsf{XY}|\mathsf{HD}]$ with $\mathsf{Y} \perp\!\!\!\perp_{\mathbf{T}} \mathsf{X}|\mathsf{HD}$, and suppose that the random variable $\mathsf{Y}$ 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 $l(h, d) = \sum_{y\in\mathbb{R}} y\mathbf{T}[\mathsf{Y}|\mathsf{HD}]_{h,d}^y$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{P}_\alpha[\mathsf{XDY}|\mathsf{H}]}[\mathsf{Y}] &= \sum_{x\in X}\sum_{d\in D}\sum_{y\in Y} y\,(\mathbf{T}[\mathsf{X}|\mathsf{H}] \rightrightarrows \mathbf{S}_\alpha[\mathsf{D}|\mathsf{X}] \rightrightarrows \mathbf{T}[\mathsf{Y}|\mathsf{XHD}])_h^{xdy} && (60)\\
&= \sum_{x\in X}\sum_{d\in D}\sum_{y\in Y} \mathbf{T}[\mathsf{X}|\mathsf{H}]_h^x \mathbf{S}_\alpha[\mathsf{D}|\mathsf{X}]_x^d \mathbf{T}[\mathsf{Y}|\mathsf{HD}]_{h,d}^y && (61)\\
&= \sum_{x\in X}\sum_{d\in D} \mathbf{T}[\mathsf{X}|\mathsf{H}]_h^x (S_\alpha)_x^d l(h, d) && (62)\\
&= R(h, \mathbf{S}_\alpha) && (63)
\end{aligned}$$

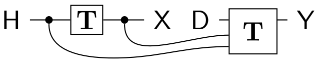
That is, if we are given a see-do model where we interpret $\mathbf{T}[\mathsf{X}|\mathsf{H}]$ as a statistical experiment and $\mathsf{Y}$ as a reversed utility, the expectation of the utility under the strategy forecast given in equation 59 is the risk of that strategy under hypothesis $h$.

## 3.2 Combs

The see-do model $\mathbf{T}[\mathsf{XY}|\mathsf{HD}]$ is known as 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.

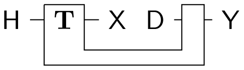
A comb is a Markov kernel with a "hole" in it. We combine the see-do model with a strategy by putting the strategy "in the middle" of the see-do model (Equation 59), rather than attaching it to one end. While it is not a well-formed diagram in the language described in this paper, we can visualise combs as Markov kernels with holes:

$\mathbf{T}[\mathsf{XY}|\mathsf{HD}] =$



(64)

$=$



(65)

We can take any strategy $\mathbf{S}_\alpha[\mathsf{D}|\mathsf{X}]$ and drop it into the "hole" in 65 (as described in Equation 59) to get a forecast of the outcome of that strategy.

Dawid (2020) has described his 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.

Lattimore and Rohde (2019a) and Lattimore and Rohde (2019b) describe a novel approach to causal inference: rather than consider "one" causal model, they consider a pair of models; an observational and interventional model that share parameters.

# 4 Causal Bayesian Networks

In the presentation of Pearl (2009), a Causal Bayesian Network posits an observational probability distribution such as $P(X, Y)$, and a set of interventional distributions, for example $\{\mathbb{P}_h(X, Y|do(X = x))\}_{x\in X, h\in H}$. Here we use notation similar to typical notation used for Causal Bayesian Networks and don't intend for these to necessarily be elements of any modelling context. For simplicity, we will consider a Causal Bayesian Network with only hard interventions on a single variable, e.g. interventions only of the form $do(\mathsf{X} = x)$.

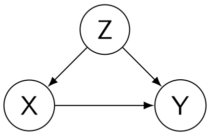
First we will offer some commentary

We can consider this an instance of a see-do model. To do so consistently within a modelling context $\mathcal{M}$, we need to distinguish observation and intervention variables - let the former retain the labels $\mathsf{X}, \mathsf{Y}$ and call the latter $\mathsf{X}', \mathsf{Y}'$.

Let $D = \{do(\mathsf{X} = x)\}_{x\in X}$. Then a Causal Bayesian Network can be considered a see-do model $\mathbf{T}[\mathsf{XYX'Y'}|\mathsf{HD}]$ by identifying $\mathbf{T}[\mathsf{XY}|\mathsf{H}]_h := \mathbb{P}_h(X, Y)$ and $\mathbf{T}[\mathsf{X'Y'}|\mathsf{HD}]_{h,do(\mathsf{X}=x)} := P_h(X, Y|do(X = x))$.

We need to rename the consequence variables because otherwise we would have $\mathbf{T}[\mathsf{XYXY}|\mathsf{HD}]$ and the two $\mathsf{X}$'s and the two $\mathsf{Y}$'s would be deterministically equal by the "identical labels" rule

We can say a bit more about Causal Bayesian Networks. Suppose we have the network



Then, letting $\mathbf{T}[\mathsf{XYZ}|\mathsf{H}]$ be the observational "see" model and $\mathbf{T}[\mathsf{X'Y'Z'}|\mathsf{HD}]$ be the interventional "do" model with $D$ the set of interventions $\{do(\mathsf{X} = x)\}_{x\in X}$ where we write $x := do(\mathsf{X} = x)$ for short, then we know by the backdoor adjustment rule that $\mathbf{T}[\mathsf{X'Y'Z'}|\mathsf{HD}]^{x'yz}_{hx} \overset{krn}{=} \mathbf{T}[\mathsf{Z}|\mathsf{H}]^z_h \delta[x]^{x'} \mathbf{T}[\mathsf{Y}|\mathsf{XZH}]^y_{hx'z}$.

Let $\mathbf{U}[\mathsf{ZY}|\mathsf{XH}] = \mathbf{T}[\mathsf{Z}|\mathsf{H}] \rightrightarrows \mathbf{T}[\mathsf{Y}|\mathsf{XZH}]$, call $\mathbf{T}[\mathsf{X}|\mathsf{H}]$ the "observational strategy" and $\mathbf{D}_x[\mathsf{X}|\mathsf{D}]^{x'}_x \overset{krn}{=} \delta[x]^{x'}$ the interventional strategies for all $x \in X$. Then we have

$$\mathbf{T}[\mathsf{XYZ}|\mathsf{H}] = \mathbf{U}[\mathsf{Z}|\mathsf{H}] \rightrightarrows \mathbf{T}[\mathsf{X}|\mathsf{H}] \rightrightarrows \mathbf{U}[\mathsf{Y}|\mathsf{XHZ}] \tag{66}$$

$$\mathbf{T}[\mathsf{X'Y'Z'}|\mathsf{HD}] \overset{krn}{=} \mathbf{U}[\mathsf{Z}|\mathsf{H}] \rightrightarrows \mathbf{D}[\mathsf{X}|\mathsf{D}] \rightrightarrows \mathbf{U}[\mathsf{Y}|\mathsf{XHZ}] \tag{67}$$

So this simple example of a Causal Bayesian network is a "nested comb" where the outer comb $\mathbf{T}[\mathsf{XYZX'Y'Z'}|\mathsf{HD}]$ is the "see" and "do" models, which are themselves generated by the inner comb $\mathbf{U}[\mathsf{ZY}|\mathsf{XH}]$ with different choices $\mathbf{T}[\mathsf{X}|\mathsf{H}]$ and $\mathbf{D}[\mathsf{X}|\mathsf{D}]$ for the insert.

This is a simple example, but Jacobs et al. (2019) has used an "inner comb" representation of a general class of Causal Bayesian Networks to prove a sufficient identification condition which is itself slightly more general than the identification condition given by Tian and Pearl (2002).

# 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 $\mathsf{Y}$ Thing"), which we represent with a variable $\mathsf{Y}$. 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 $\mathsf{W}$ 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 $\mathsf{Y}$ would be" as "what would $\mathsf{Y}$ be if $\mathsf{W}$ was 0 if $\mathsf{W}$ 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 $\mathsf{Y}$. 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 $\mathsf{Y}$ and $\mathsf{Y}(W)$ are random variables and $\mathsf{W}$ 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 $\mathbf{K}[\mathsf{Y}|\mathsf{W}\mathsf{Y}(W)]_{ij_1j_2\ldots j_{|W|}} = \delta[j_i]$.

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

We allow $\mathsf{X}$ 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 $\mathsf{X}$ is a random variable) or as submodels of consequence models (in which case $\mathsf{X}$ 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 $\mathbf{K}[\mathsf{Y}|\mathsf{X}\mathsf{Y}(X)]_{ij_1j_2\ldots j_{|X|}} = j_i$ (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 $\mathsf{Y}$ would take if $\mathsf{X}$ 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 $\mathsf{Y}$ represents TYT and $\mathsf{W}$ represents the regime, then we have "for all $w$, $\mathsf{Y}(w)$ is equal to $\mathsf{Y}$ 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 $\mathsf{Y}_i$ and $\mathsf{Y}(W)$ are random variables and $\mathsf{W}_i$ could be either a state or random variables is a *parallel potential outcome submodel* if $\mathbf{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]$ exists and $\mathbf{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]_{kj_1j_2\ldots j_{|W|}} = \delta[j_k]$.

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

A parallel potential outcomes model features a sequence of $n$ “parallel” outcome variables $\mathsf{Y}_i$ and $n$ “regime proposals” $\mathsf{W}_i$, 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 $\mathsf{Y}_1$ represents the value of TYT in the actual world and $\mathsf{Y}_i$ $i \neq 1$ represents TYT under a supposed regime $\mathsf{W}_i$. 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 $\mathsf{W}_i$ is $w$ then the corresponding outcome will be almost surely equal to $\mathsf{Y}(w)$, regardless of the value of the actual regime $\mathsf{W}_1$”, 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 $\mathsf{X}_1$ is the result of the first flip of a coin that we know to be fair and $\mathsf{X}_2$ 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 $\mathsf{Y}_1$ and $\mathsf{Y}_2$ 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 $\mathsf{Y}_1, \mathsf{Y}_2, \ldots$ can be modeled by the product of a *prior* probability $\mathbb{P}[\mathsf{J}]$ with $\mathsf{J}$ 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 $\mathsf{J}$ 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 $\mathsf{W}_j$ and $\mathsf{W}_i$ are equal then $\mathsf{Y}_j$ and $\mathsf{Y}_i$ 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 $\mathsf{Y}_i$ random variables is *exchangeable* conditional on $\mathsf{Z}$ 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 $\mathsf{X}$ and some state variable $\mathsf{U}$ 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 $\mathsf{Y}$ on $(\mathbf{K}, E, F)$ with $\mathsf{Y} \circ \mathsf{X}$ on $(\mathbf{K}', E', F')$ and any state variable $\mathsf{D}$ 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 $\mathsf{X}$ and $\mathbb{P}(\mathsf{X})$ then introduce some random variable $\mathsf{Y}$ and $\mathbb{P}(\mathsf{XY})$ as if the sample space contained both $\mathsf{X}$ and $\mathsf{Y}$ 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 $\mathsf{Z}$ with $\mathsf{Y}_i$ 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 $\mathsf{Z}$.

Here that we identify $\mathsf{Z}$ 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 $\mathsf{Y}_i$ random variables is *deterministically reproducible* conditional on $\mathsf{Z}$ 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 $\mathsf{Z}$, $(\mathbf{K}, E, F)$ 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 $\mathsf{Z}$.*

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

$$\mathbf{K}[\mathsf{X}_n|\mathsf{D}_n\mathsf{Z}]^x_{dz} = \sum_{y\in Y^{|D|}} \mathbf{K}[\mathsf{X}_{[n]}|\mathsf{D}_n\mathsf{Z}\mathsf{Y}(D)]^x_{dzy}\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_{[n]}\mathsf{Z}]^y_{dz} \tag{68}$$

$$= \sum_{y\in Y^{|D|}} \prod_{i\in[n]} \delta[y_{d_i}]^{x_i}\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_n\mathsf{Z}]^y_{dz} \tag{69}$$

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

$$\begin{aligned}
\mathbf{K}[\mathsf{X}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]^x_{dz} &= \sum_{y\in Y^{|D|}} \prod_{i\in Q_d} \delta[y_{d_i}]^{x_i} \mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_{[n]}\mathsf{Z}]^y_{dz} && (70)\\
&= \sum_{y'\in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}(R_d)\mathsf{Y}(R_d^C)|\mathsf{D}_{Q_d}\mathsf{D}_{Q_d^C}\mathsf{Z}]^{x_{Q_d}y'}_{d_{Q_d}d^C_{Q_d}z} && (71)\\
&= \sum_{y'\in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}_{R_d}\mathsf{Y}_{R_d^C}|\mathsf{D}_{Q_d}\mathsf{D}_{Q_d^C}\mathsf{Z}]^{x_{Q_d}y'}_{dz} && (72)\\
&= \sum_{y'\in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]^{x_{Q_d}y'}_{dz} \qquad \text{(using exchangeability)}\\
& && (73)
\end{aligned}$$

Note that

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

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

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

$$\mathbb{P}[\mathsf{XDY}]^{ijk} = \mathbb{P}[\mathsf{X}]^i \mathbb{P}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{P}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \tag{74}$$

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

$$\mathbf{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^l_{kj} \tag{75}$$

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

We then have

$$\begin{aligned}
\mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj} &= \mathbf{T}\underline{\mathsf{X}}'^k_{hj} && (76)\\
&= \sum_l \mathbf{T}^{kl}_{hj} && (77)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^k && (78)\\
&= \mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj'} && (79)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}_{h,\mathbf{S}}[\mathsf{XDY}]^{ijk} &= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} && (80)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} && (81)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]^j_i \mathbf{T}[\mathsf{X}'\mathsf{Y}'|\mathsf{H}'\mathsf{D}']^{ik}_{hj} && (82)\\
&= \mathbf{S}^j_i \mathbf{T}[\mathsf{X}'\mathsf{Y}'|\mathsf{H}'\mathsf{D}']^{ik}_{hj} && (83)\\
&= \mathbf{S}^j_i \mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^i_{hj} \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} && (84)\\
&= \mathbf{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbf{S}^j_i \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} && (85)
\end{aligned}$$

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

$$\mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i = \mathbb{P}_{h,\mathbf{Q}}[\mathsf{X}]^i \tag{86}$$

$$\mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} = \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \text{ for some version of } \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}] \tag{87}$$

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

$$\mathbf{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbf{Q}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}]^l_{kj} \tag{88}$$

Then by substitution of $\mathbf{Q}$ for $\mathbf{S}$ in Equation 80 and working through the same steps

$$\mathbb{P}_{h,\mathbf{S}}[\mathsf{XDY}]^{ijk} = \mathbf{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbf{Q}^j_i \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} \tag{89}$$

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

$$\mathsf{I}_{K\cdot\cdot} := \mathsf{I}_K \setminus \mathsf{I}_L \setminus \mathsf{I}_J \tag{90}$$

$$\mathsf{I}_{KL\cdot} := \mathsf{I}_K \cap \mathsf{I}_L \setminus \mathsf{I}_J \tag{91}$$

$$\mathsf{I}_{K\cdot J} := \mathsf{I}_K \cap \mathsf{I}_J \setminus \mathsf{I}_L \tag{92}$$

$$\mathsf{I}_{KLJ} := \mathsf{I}_K \cap \mathsf{I}_L \cap \mathsf{I}_J \tag{93}$$

$$\mathsf{I}_{\cdot L\cdot} := \mathsf{I}_L \setminus \mathsf{I}_K \setminus \mathsf{I}_J \tag{94}$$

$$\mathsf{I}_{\cdot LJ} := \mathsf{I}_L \cap \mathsf{I}_J \setminus \mathsf{I}_K \tag{95}$$

$$\mathsf{I}_{\cdot\cdot J} := \mathsf{I}_J \setminus \mathsf{I}_K \setminus \mathsf{I}_L \tag{96}$$

$$\mathsf{O}_{K\cdot\cdot} := \mathsf{O}_K \setminus \mathsf{I}_N \setminus \mathsf{I}_J \tag{97}$$

$$\mathsf{O}_{KL\cdot} := \mathsf{O}_K \cap \mathsf{I}_L \setminus \mathsf{I}_J \tag{98}$$

$$\mathsf{O}_{K\cdot J} := \mathsf{O}_K \cap \mathsf{I}_J \setminus \mathsf{I}_L \tag{99}$$

$$\mathsf{O}_{KLJ} := \mathsf{O}_K \cap \mathsf{I}_L \cap \mathsf{I}_J \tag{100}$$

$$\mathsf{O}_{L\cdot} := \mathsf{O}_L \setminus \mathsf{I}_J \tag{101}$$

$$\mathsf{O}_{LJ} := \mathsf{O}_L \cap \mathsf{I}_J \tag{102}$$

Also define

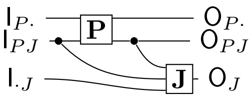
$$(\mathbf{P}, \mathsf{I}_P, \mathsf{O}_P) := \mathbf{K} \rightrightarrows \mathbf{L} \tag{103}$$

$$(\mathbf{Q}, \mathsf{I}_Q, \mathsf{O}_Q) := \mathbf{L} \rightrightarrows \mathbf{J} \tag{104}$$

Then

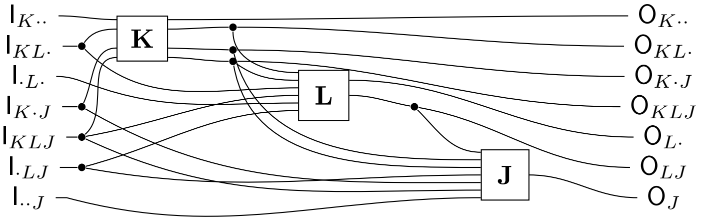
$$(\mathbf{K} \rightrightarrows \mathbf{L}) \rightrightarrows \mathbf{J} = \mathbf{P} \rightrightarrows \mathbf{J} \tag{105}$$

$=$



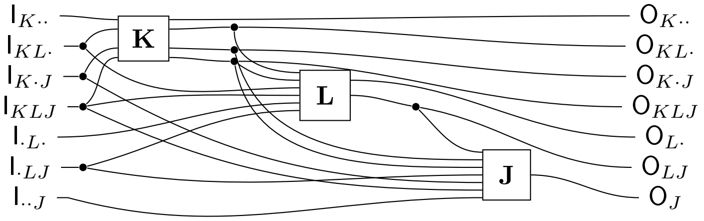
(106)

$=$



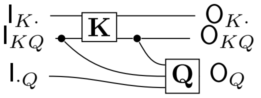
(107)

$\overset{perm}{=}$



(108)

$=$



(109)

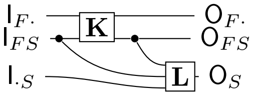
$$= \mathbf{K} \rightrightarrows (\mathbf{L} \rightrightarrows \mathbf{J}) \tag{110}$$

# 8 Appendix: String Diagram Examples

Recall the definition of *connection*:

**Definition 8.1** (Connection)**.**

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



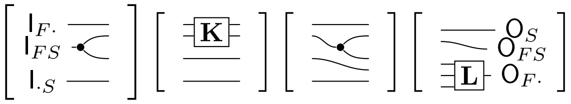
(111)

$$:= \mathbf{J} \tag{112}$$

$$\mathbf{J}^{zxw}_{yqr} = \mathbf{K}^{zx}_{yq}\mathbf{L}^{w}_{xqr} \tag{113}$$

Equation 111 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:

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



(114)

(115)

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

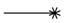
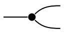
Note that this is a more general definition than the one used in the main paper; the version in the main paper is the restriction of this definition to finite sets.

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

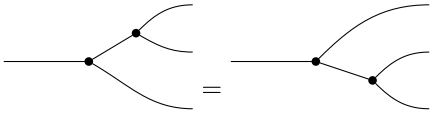
$$\delta(x, A) = \begin{cases} 1 & x \in A \\ 0 & \text{otherwise} \end{cases} \tag{116}$$

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

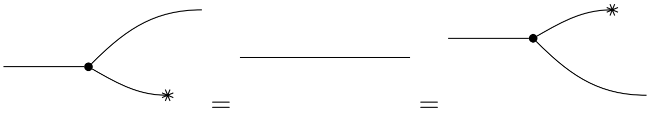
**Definition 9.1.** A Markov category $C$ is a symmetric monoidal category in which every object $X \in C$ is equipped with a commutative comonoid structure given by a comultiplication $\text{copy}_X : X \to X \otimes X$ and a counit $\text{del}_X : X \to I$, depicted in string diagrams as

$\text{del}_X :=$  $\text{copy}_X$ $:=$  (117)

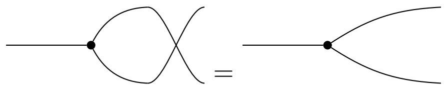
and satisfying the commutative comonoid equations



(118)

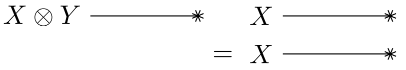


(119)

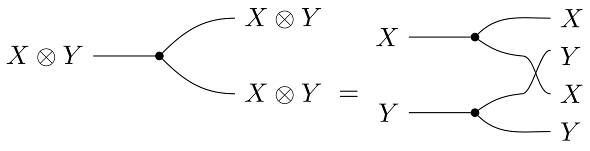


(120)

as well as compatibility with the monoidal structure

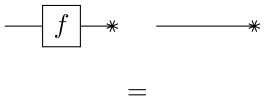


(121)



(122)

and the naturality of *del*, which means that



(123)

for every morphism $f$.

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

$$\mathbf{KL} : \mathsf{X} \to \mathit{\Delta}(\mathsf{Z}) \tag{124}$$

$$\text{defined by } (\mathbf{KL})(x, A) = \int_Y \mathbf{L}(y, A)\mathbf{K}(x, dy) \tag{125}$$

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

$$\mathbf{K} \otimes \mathbf{L} : (\mathsf{X}, \mathsf{W}) \to \mathit{\Delta}(\mathsf{Y}, \mathsf{Z}) \tag{126}$$

$$\text{defined by } \mathbf{K} \otimes \mathbf{L}(x, w, A \times B) = \mathbf{K}(x, A)\mathbf{L}(w, B) \tag{127}$$

The identity map is

$$\mathrm{Id}_{\mathsf{X}} : \mathsf{X} \to \mathit{\Delta}(\mathsf{X}) \tag{128}$$

$$\text{defined by}(\mathrm{Id}_X)(x, A) = \delta(x, A) \tag{129}$$

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

The swap map is given by

$$\mathrm{swap}_{\mathsf{X},\mathsf{Y}} : (\mathsf{X},\mathsf{Y}) \to \varDelta(\mathsf{Y},\mathsf{X}) \tag{130}$$

$$\text{defined by}(\mathrm{swap}_{\mathsf{X},\mathsf{Y}})(x, y, A \times B) = \delta(x, B)\delta(y, A) \tag{131}$$

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

The copy map is given by

$$\mathrm{copy}_{\mathsf{X}} : \mathsf{X} \to \varDelta(\mathsf{X},\mathsf{X}) \tag{132}$$

$$\text{defined by}(\mathrm{copy}_{X})(x, A \times B) = \delta_x(A)\delta_x(B) \tag{133}$$

and the erase map by

$$\mathrm{del}_{\mathsf{X}} : \mathsf{X} \to \varDelta(*) \tag{134}$$

$$\text{defined by}(\mathrm{del}_{X})(x, A) = \delta(*, A) \tag{135}$$

$$\tag{136}$$

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 118 to 123 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.

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.

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.

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

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

## Appendix: