Understanding Causal Primitives Using Modular Probability

David Johnston

September 28, 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 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

[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.

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 Probability with connectable submodels

Throughout this paper, we will work with a restricted probability theory in which any measurable set X to be a finite set. This is both because it makes explanations simpler and because it is easy to show that submodels exist in this setting (Lemma 2.7). Many of the proofs in this paper (with the exception of the one just mentioned) can likely be specialised to more general probility

theories due to our use of string diagrams developed for Markov categories.

The standard method of constructing probability models introduces a probability space $(\mathbb{P}, (\Omega, \mathcal{F}))$ with Ω a sample space, $\mathcal{F}$ a σ-algebra on Ω and $\mathbb{P}$ a probability measure on $(\Omega, \mathcal{F})$. Random variables are defined by measurable functions on Ω and are given names in sans-serif like X. A probability distribution $\mathbb{P}^{\mathsf{XYZ}}$ is "the joint distribution of X, Y and Z under $\mathbb{P}$" where X, Y and Z are associated with random variables on Ω and is given by the pushforward of the function $\omega \mapsto (\mathsf{X}(\omega), \mathsf{Y}(\omega), \mathsf{Z}(\omega))$. Unless otherwise stated, a random variable named X will take values in the space X (note the serif font).

To help us to construct causal models, we need an alternative method for building probability models. The motivation for doing this is to make it easy to represent a certain type of operation common in causal models. It is well-known that causal models make use of operations that are not standard in probability theory. For example, in the causal graphical model framework, given $\mathbb{P}^{\mathsf{XYZ}}$, if Z blocks a backdoor path between X and Y then the backdoor adjustment formula allows us to define a new probability space with measure $\mathbb{P}_x$ via "truncated factorisation" (Pearl, 2009, page 24):

$$\mathbb{P}_x^{\mathsf{YZ}}(y, z) := \mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}(y|x, z)\mathbb{P}^{\mathsf{Z}}(z) \tag{1}$$

The standard theory of probability does not assign any special siginificance to the expression on the right side of Equation 1. At the same time, the notation we have chosen for the left side of Equation 1 implicitly claims that $\mathbb{P}_x^{\mathsf{YZ}}$ is a distribution over the same variables Y and Z as the original $\mathbb{P}^{\mathsf{XYZ}}$. One way we can make sense of this is if $\mathbb{P}_x$ is defined on the same sample space as the original $\mathbb{P}$, and Y and Z are the same measurable functions on Ω. However, we need to be careful that Equation 1 is not therefore a contradiction. For example, if $\mathsf{X} = \mathsf{Z}$ (as in, X and Z are *the same function on* Ω) then there will generally be no measure $\mathbb{P}_x$ such that $\mathbb{P}_x^{\mathsf{YZ}}(y, z)$ agrees with Equation 1.

We want to be able to define operations like Equation 1, and we want to keep the idea that the measure that results from this operation is a joint distribution over the "same variables". We do this by distinguishing *variable names* from the functions used to represent the variables on a particular sample space. In our framework, variables are measurable functions *with names*; most measurable functions will not be given a name, and so are not variables. We note that this is implicitly accomplished by the Structural Causal Model framework, in which "intervention" takes a variable named X associated with a function $f : \Omega \to X$ and replaces this with a different function $g : \Omega \to X$ associated with the same variable X. This operation cannot be defined with an "anonymous function" because there is no name enabling a statement like "this new function g points to the same thing the old function f did". Consequently, "variable names" are an integral part of Structural Causal Models.

2.1 Markov categories

The basic elements we will work with are finite sets and Markov kernels. A Markov kernel $\mathbf{K} : X \to \Delta(Y)$ is a map from X to probability distributions on Y. We can represent it concretely by the elements $(\mathbf{K}_x^y)_{x \in X, y \in Y}$. An element K_x^y represents the probability of $y \in Y$ given the argument $x \in X$. In general, an argument w appearing as a superscript can be read as "the probability of w" and an argument v appearing as a subscript can be read as "given v". Note that we do *not* use Einstein summation in any expressions in this paper – all sums will be written out explicitly.

A Markov kernel must have the following properties:

$$0 \leq K_x^y \leq 1 \qquad \forall x, y \tag{2}$$

$$\sum_{y \in Y} K_x^y = 1 \qquad \forall x \tag{3}$$

A probability distribution is a Markov kernel $\mathbf{P} : \{*\} \to \Delta(Y)$ where $\{*\}$ is a one-element set. Such a Markov kernel can be represented as a matrix with one row, i.e. a column vector.

We define two particular Markov kernels that play a special role. The erase map $\top : X \to \{*\}$ is represented by the matrix $\top_x = 1$ for all $x \in X$. It maps every element of X to the unique probability distribution on $\{*\}$, which gives probability 1 to $*$, the only element of the set; we can think of this as forgetting the input.

The copy map $\curlyvee : X \to \Delta(X \times X)$ is the Markov kernel represented by the matrix $\curlyvee_x^{x', x''} := [\![x = x']\!][\![x = x'']\!]$, where the iverson bracket $[\![\cdot]\!]$ evaluates to 1 if $\cdot$ is true and 0 otherwise. We can think of the copy map as taking an element x and outputting a joint distribution of two "variables" that are deterministically equal to x.

A swap map $\between : X \times Y \to \Delta(Y \times X)$ is the Markov kernel represented by the matrix $\between_{x,y}^{y',x'} := [\![x = x']\!][\![y = y']\!]$. It takes two inputs and returns a joint distribution on two "variables" deterministically equal to the swapped inputs.

The category with finite sets as objects, Markov kernels as morphisms, matrix products as the composition operation, $\top$ as the counit and $\curlyvee$ as the comultiplication forms the category FinStoch. This is not a category theory paper, but the fact that we are working in this category has a practical consequence. FinStoch is a *Markov category*, as defined by Fritz (2020) and discussed earlier by Cho and Jacobs (2019); Fong (2013). All Markov categories share a formal system of "string diagrams" such that a valid derivation using the diagrammatic notation corresponds to a valid theorem in the category. Markov categories include categories more general than ours, such as the category with general measurable sets as objects and Markov kernels as morphisms.

2.2 Graphical notation for Markov categories

In an appendix, state the axioms of Markov categories and maybe a short tutorial on reading diagrams

We represent a Markov kernel as a box and a probability distribution as a triangle:

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

$$\mathbb{P} := \quad \triangleleft\!\boxed{\mathbf{K}}- \tag{5}$$

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

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

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)). They are to be read as a series of products of Markov kernels. For some special Markov kernels, we can replace the generic "box" of a Markov kernel with a special diagrammatic element that is visually suggestive of what they accomplish.

The Cartesian product $X \times Y := \{(x, y) | x \in X, y \in Y\}$.

Given kernels $\mathbf{K} : W \to Y$ and $\mathbf{L} : X \to Z$, the tensor product $\mathbf{K} \otimes \mathbf{L} : W \times X \to \Delta(Y \times Z)$ is defined by $(\mathbf{K} \otimes \mathbf{L})_{(w,x)}^{(y,z)} := K_w^y L_x^z$.

The tensor product is represeted by parallel juxtaposition:

$$\mathbf{K} \otimes \mathbf{L} := \quad \begin{array}{l} -\boxed{\mathbf{K}}- \\ -\boxed{\mathbf{L}}- \end{array} \tag{7}$$

The identity map is a bare line:

$$\mathrm{Id} := \quad - \tag{8}$$

The stopper is a fuse:

$$\top := \quad \longrightarrow\!\!* \tag{9}$$

The splitter is a fork:

$$\curlyvee := \quad -\!\!\bullet\!\!< \tag{10}$$

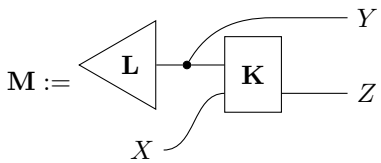
The swap map swaps wires:

$\mathsf{X} :=$  (11)

We will use the graphical notation for derivations, but because it is quite unfamiliar we will also include translations to more familiar notation.

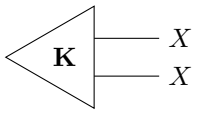
2.3 Revisiting truncated factorisation

Recall the original problem of defining operations like Equation 1. We can seemingly do this quite easily using the tools of the FinStoch category. Note that $P^{\mathsf{Y}|\mathsf{XZ}}$ must be represented by a Markov kernel $\mathbf{K} : X \times Z \to \Delta(Y)$ and $\mathbb{P}^{\mathsf{Z}}$ by a Markov kernel $\mathbf{L} \in \Delta(Z)$ (we will explain later why we explicitly distinguish Markov kernels from conditional probabilities). Then it seems that we can define a Markov kernel $\mathbf{M} : X \to \Delta(X \times Z)$ representing $x \mapsto \mathbb{P}_x^{\mathsf{YZ}}(y, z)$ by

$\mathbf{M} :=$ 

(12)

There is a problem, however: in the diagram above X, Y and Z refer to the *sets in which variables take values*, which are not an adequate substitute for variable labels. For example, if X_1 and X_2 both take values in X, then a Markov kernel $\mathbf{K}$ representing some $\mathbb{P}^{\mathsf{X}_1\mathsf{X}_2}$ could be drawn

$\mathbf{K} :=$ 

(13)

However the variable associated with each wire is different. Our solution to this is to give labels to sets. A labeled set is a set together with a label, such as (X_1, X), which can be read as the set of statements "$\mathsf{X}_1 = x$" for each $x \in X$. In fact, we allow sets to have multiple synonymous labels $(\{\mathsf{X}_1, \mathsf{Q}\}, X)$, which can be read as stipulating that X_1 and Q refer to the same thing. In general, any two labels that appear in the same label set are called *synonyms*.

Let N_{X} refer to the set of labels of X. Any collection of labeled sets must live in a "labelspace" that prevents us from making inadmissible label assignments. We require the following axioms to hold for labeled sets occupying the same labelspace:

1. **Uniqueness of labels:** Given two labeled sets (N_1, X) and (N_2, Y), if there exists some $\mathsf{W} \in N_1$ and $\mathsf{W} \in N_2$ then $N_1 = N_2$ and $X = Y$

2. **Empty label:** There is a unique empty label $*$ which is always associated with a 1-element set $\{*\}$

Due to Axiom 1, we can unambiguously refer to a labeled set (N_1, X) by any of its labels $\mathsf{X} \in N_1$.

We define a *sequence* of labeled sets

- **Sequence of labeled sets:** Given (N_{X}, X) and (N_{Y}, Y) with $\mathsf{X} \in N_{\mathsf{X}}$ and $\mathsf{Y} \in N_{\mathsf{Y}}$, the sequence (X, Y) is the labeled set $(N, X \times Y)$ with $(\mathsf{X}, \mathsf{Y}) \in N$

Note that (X, Y) is also a label. (X, Y) may have synonyms such as Z. We say a label U is *atomic* if none of its synonyms are sequences of labels.

For some W, we use $m_{\mathsf{W}}(\mathsf{Z})$ to refer to "the number of times W appears in Z". We define this as follows:

- $m_{\mathsf{W}}(\mathsf{Z}) = 1$ if W is a synonym of Z
- $m_{\mathsf{W}}(\mathsf{Z}) = 0$ if Z is atomic and W is not a synonym of Z
- Otherwise, if Z is synonymous with (X, Y) then $m_{\mathsf{W}}(\mathsf{Z}) = \arg\max_{\mathsf{X}' \in N_{\mathsf{X}}} m_{\mathsf{W}}(\mathsf{X}') + \arg\max_{\mathsf{Y}' \in N_{\mathsf{Y}}} m_{\mathsf{W}}(\mathsf{Y}')$

It's not obvious that this is unique

We say X is in Y or $\mathsf{X} \in \mathsf{Y}$ if $m_{\mathsf{X}}(\mathsf{Y}) > 0$. We say that $*$ is in every sequence.

With this in mind, we define the following operations on sequences of labels

1. **Difference of labels:** Given X, Y, the difference $\mathsf{X} \setminus \mathsf{Y}$ is a label sequence Z such that for any label X_i, $m_{\mathsf{X}_i}(\mathsf{Z}) = \max(0, m_{\mathsf{X}_i}(\mathsf{X}) - m_{\mathsf{X}_i}(\mathsf{Y}))$
2. **Intersection of labels:** Given X, Y, the intersection $\mathsf{X} \cap \mathsf{Y}$ is a label sequence Z such that for any label X_i, $m_{\mathsf{X}_i}(\mathsf{Z}) = \min(m_{\mathsf{X}_i}(\mathsf{X}), m_{\mathsf{X}_i}(\mathsf{Y}))$

These definitions are non-unique in that they do not define the order of Z. This doesn't cause a problem because we work with labeled Markov kernels that we define to be equivalent if one can be obtained from the other by permuting its labels and applying the corresponding swap maps. We use the convention that $*$ is synonymous with a sequence under which every label has multiplicity 0.

2.4 Labeled Markov kernels

LabeledFinStoch is actually a different category to FinStoch and I have to show it satisfies the axioms of a Markov category

A labeled Markov kernel $\mathbf{K} : \mathsf{A} \to \Delta(\mathsf{B})$ is a Markov kernel that maps between labeled sets. If A takes values in X and B takes values in Y, then $\mathbf{K}$ is the underlying Markov kernel $\mathbf{K}' : X \to \Delta(Y)$ along with the *domain label* A and *codomain label* B. The labels A and B can be replaced by any synonyms of A and B.

A labeled probability distribution $\mathbb{P} \in \Delta(Y)$ comes with a codomain label (B) only.

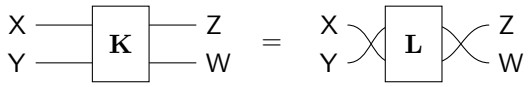
Graphically, we annotate the wires of a labeled kernel with the corresponding labels. If $\mathbf{K} : (\mathsf{A}_1, \mathsf{A}_2) \to \Delta(\mathsf{B}_1, \mathsf{B}_2)$, it is represented:

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

We can also use single wires to represent sequences of labeled sets:

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

We say two labeled kernels $\mathbf{K}$ and $\mathbf{L}$ are equivalent if one can be obtained from the other by a permutation of labels and application of the corresponding swap maps. For example, if we have $\mathbf{K} : (\mathsf{X}, \mathsf{Y}) \to \Delta(\mathsf{Z}, \mathsf{W})$ and $\mathbf{L} : (\mathsf{Y}, \mathsf{X}) \to \Delta(\mathsf{W}, \mathsf{Z})$, then $\mathbf{K}$ and $\mathbf{L}$ are equivalent, written $\mathbf{K} \overset{perm}{=} \mathbf{L}$ if



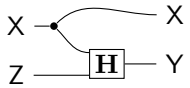
(16)

A labeled Markov kernel must satisfy the following axiom. This axiom is what makes the use of labeled Markov kernels and labeled sets different to the use of ordinary Markov kernels and sets.

1. For any labeled kernel $\mathbf{K} : \mathsf{A} \to \Delta(\mathsf{B})$, there must be a valid diagrammatic representation in which all instances of the same label are connected by a path consisting of wires and copy maps only.

For example, given $\mathbf{K} : (\mathsf{X}, \mathsf{Y}) \to \Delta(\mathsf{X}, \mathsf{Z})$, we require that there exist some $\mathbf{H} : X \times Z \to \Delta(Y)$ such that

$$\mathbf{K} = \tag{17}$$

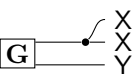


$$\iff \tag{18}$$

$$\mathbf{K}^{x'y}_{xz} = [\![x = x']\!]\mathbf{H}^{y}_{xz} \tag{19}$$

Note the path connecting the two instances of X goes only through a copy map. For a second example, given $\mathbf{L} : \mathsf{Z} \to \Delta(\mathsf{X}, \mathsf{X}, \mathsf{Y})$, we require that there exist some $\mathbf{G} : \mathsf{Z} \to \Delta(\mathsf{X}, \mathsf{Y})$ such that

$$\mathbf{L} = \tag{20}$$



$$\iff \tag{21}$$

$$\mathbf{L}^{xx'y}_{z} = [\![x = x']\!]\mathbf{G}^{xy}_{z} \tag{22}$$

In short, *instances of the same label must be deterministically equal.* Note that this axiom rules out the existence of any labeled kernels with multiple copies of the same label as inputs.

The *connection* operation, defined in the next section, can compactly represent both cases above. If X is some label that appears twice in the output of $\mathbf{L}$ or once in the input and once in the output, then there must be some $\mathbf{M}$ such that

$$\mathbf{L} = \mathbf{M} \rightrightarrows \mathbf{I} \tag{23}$$

I need to make this an axiom

2.5 Connection

Connection is an associative operation $\rightrightarrows$ that "joins" two labeled Markov kernels where the labels can be matched and preserves unmatched inputs and outputs. A key property of connection is that, if both input Markov kernels satisfy Axiom 1, then the output also satisfies axiom 1. One can think of this operation like connecting two lego bricks of different sizes – we connect all the parts that will fit together, and all the connection points that don't fit are left available.

Given two labeled Markov kernels $\mathbf{F} : \mathsf{I}_F \to \Delta(\mathsf{O}_F)$ and $\mathbf{S} : \mathsf{I}_S \to \Delta(\mathsf{O}_S)$, make the following label identifications:

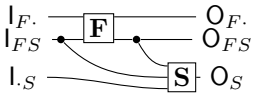
$$\mathsf{O}_{F\cdot} := \mathsf{O}_F \setminus \mathsf{I}_S \qquad \text{Labels only in the output of } \mathbf{F} \tag{24}$$
$$\mathsf{O}_{FS} := \mathsf{O}_F \cap \mathsf{I}_S \qquad \text{Labels in the output of both} \tag{25}$$
$$\mathsf{I}_{F\cdot} := \mathsf{I}_F \setminus \mathsf{I}_S \qquad \text{Labels only in the input } \mathbf{F} \tag{26}$$
$$\mathsf{I}_{FS} := \mathsf{I}_F \cap \mathsf{I}_S \qquad \text{Labels in the input of both} \tag{27}$$
$$\mathsf{I}_{\cdot S} := \mathsf{I}_S \setminus \mathsf{I}_F \qquad \text{Labels only in the input of } \mathbf{S} \tag{28}$$
$$\mathsf{O}_{I_F O_S *} := \mathsf{O}_S \cap \mathsf{I}_F \setminus \mathsf{I}_S \qquad \text{Input of } \mathbf{F} \text{ and the output only of } \mathbf{S} \tag{29}$$
$$\mathsf{O}_{O_F O_S *} := \mathsf{O}_F \cap \mathsf{O}_S \setminus \mathsf{I}_S \qquad \text{Output of } \mathbf{F} \text{ and the output only of } \mathbf{S} \tag{30}$$
$$\tag{31}$$

$\mathbf{F}$ can be connected to $\mathbf{S}$ iff $\mathsf{O}_{I_F O_S *}$ synonymous with $*$ and $\mathsf{O}_{O_F O_S *}$ is also synonymous with $*$. The reason for this is that, in general, if these sets were non-empty then we would not have a way to connect $\mathbf{F}$ and $\mathbf{S}$ without violating axiom 1.

Definition 2.1 (connection)**.** Consider a labeled Markov kernel $\mathbf{F} : \mathsf{I}_F \to \Delta(\mathsf{O}_F)$ which can be connected to $\mathbf{S} : \mathsf{I}_S \to \Delta(\mathsf{O}_S)$. Because they can be conected, we can write $\mathbf{F} : (\mathsf{I}_{F\cdot}, \mathsf{I}_{FS}) \to \Delta(\mathsf{O}_{F\cdot}, \mathsf{O}_{FS}))$ and $\mathbf{S} : (\mathsf{I}_{FS}, \mathsf{I}_{\cdot S}) \to \Delta(\mathsf{O}_S)$.

Then Equations 113 and 115 are equivalent definitions of extension:

$$\mathbf{K} \rightrightarrows \mathbf{L} := \quad (32)$$



$$:= \mathbf{J} \quad (33)$$

$$\mathbf{J}^{zxw}_{yqr} = \mathbf{F}^{zx}_{yq}\mathbf{S}^{w}_{xqr} \quad (34)$$

Note that there are no sums in Equation 113, this is simply a product of matrix elements.

Lemma 2.2 (Connection is associative up to permuation of labels)**.** *Given labeled Markov kernels* $\mathbf{K} : \mathsf{I}_K \to \Delta(\mathsf{O}_K)$, $\mathbf{L} : \mathsf{I}_L \to \Delta\mathsf{O}_L)$ *and* $\mathbf{J} : \mathsf{I}_J \to \Delta(\mathsf{O}_J)$,

$$(\mathbf{K} \rightrightarrows \mathbf{L}) \rightrightarrows \mathbf{J} \overset{perm}{=} \mathbf{K} \rightrightarrows (\mathbf{L} \rightrightarrows \mathbf{J}) \quad (35)$$

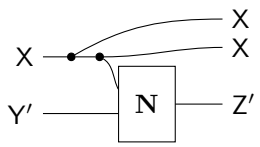
Proof. Proven in Appendix 7 □

Lemma 2.3 (Identity maps commute one way with connection)**.** *Consider the identity map on some labeled set* $\mathbf{I} : \mathsf{X} \to \Delta(\mathsf{X})$ *(note that by Equation 17 the identity map is the* only *kernel with this signature). For any* $\mathbf{M} : \mathsf{Y} \to \Delta(\mathsf{Z})$*, either a copy of* X *appears in the output but not the input and* $\mathbf{I} \rightrightarrows \mathbf{M}$ *is undefined, or* $\mathbf{I} \rightrightarrows \mathbf{M} = \mathbf{M} \rightrightarrows \mathbf{I}$*.*

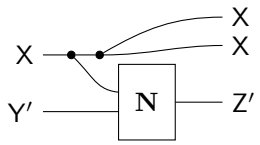
Proof. Consider the identity map on some labeled set $\mathbf{I} : \mathsf{X} \to \Delta(\mathsf{X})$ (note that by Equation 17 the identity map is the *only* kernel with this signature). Note that for any $\mathbf{M} : \mathsf{Y} \to \Delta(\mathsf{Z})$, either a copy of X appears in the output but not the input, in which case $\mathbf{I} \rightrightarrows \mathbf{M}$ is undefined, or we have one of the following cases:

If X is in Y and Z, then there must be some $\mathbf{N}$ such that Equation 17 holds. Defining $\mathsf{Y}' = \mathsf{Y} \setminus \mathsf{X}$ and $\mathsf{Z}' = \mathsf{Z} \setminus \mathsf{X}$:

$$\mathbf{I} \rightrightarrows \mathbf{M} = \quad (36)$$



$$= \quad (37)$$

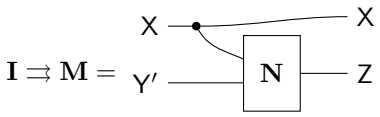


$$= \mathbf{M} \rightrightarrows \mathbf{I} \quad (38)$$

Commutativity of copy map in appendix

If X is in Y only, defining $\mathsf{Y}' = \mathsf{Y} \setminus \mathsf{X}$:

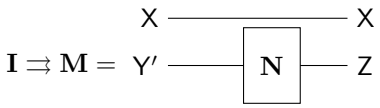
$$\mathbf{I} \rightrightarrows \mathbf{M} = \tag{39}$$



$$= \mathbf{M} \rightrightarrows \mathbf{I} \tag{40}$$

If X is in neither Y nor Z, then

$$\mathbf{I} \rightrightarrows \mathbf{M} = \tag{41}$$



$$= \mathbf{M} \rightrightarrows \mathbf{I} \tag{42}$$

□

Theorem 2.4 (Connection is compatible with axiom 1)**.** *Given* $\mathbf{K} : \mathsf{A} \to \Delta(\mathsf{B})$ *and* $\mathbf{L} : \mathsf{C} \to \Delta(\mathsf{D})$*, let* $\mathbf{J} = \mathbf{K} \rightrightarrows \mathbf{L}$*. Then* $\mathbf{J}$ *satisfies axiom 1.*

Proof. By inspecting the definition of $\rightrightarrows$ (Equation 113), we can see that no labels from either of the inputs are increased in multiplicity. We need to verify that if either of the inputs has a label with multiplicity > 1, then the result of the extension still satisfies axiom 1.

Consider any label X that appears in both the input and output of $\mathbf{K}$, or twice in the output of $\mathbf{K}$. Then Equation 23 implies that there exists some $\mathbf{H}$ such that $\mathbf{K} = \mathbf{H} \rightrightarrows \mathbf{I}$.

Thus

$$\mathbf{K} \rightrightarrows \mathbf{L} = (\mathbf{H} \rightrightarrows \mathbf{I}) \rightrightarrows \mathbf{L} \tag{43}$$

$$\overset{perm}{=} \mathbf{H} \rightrightarrows (\mathbf{I} \rightrightarrows \mathbf{L}) \tag{44}$$

$$= \mathbf{H} \rightrightarrows (\mathbf{L} \rightrightarrows \mathbf{I}) \tag{45}$$

$$\overset{perm}{=} (\mathbf{H} \rightrightarrows \mathbf{L}) \rightrightarrows \mathbf{I} \tag{46}$$

Which implies that $\mathbf{J}$ satistfies Equation 23 for Y.

Consider Z that appears in the input and output of $\mathbf{L}$ or twice in the output of $\mathbf{L}$. Then there exists some $\mathbf{G}$ such that

$$\mathbf{K} \rightrightarrows \mathbf{L} = \mathbf{K} \rightrightarrows (\mathbf{G} \rightrightarrows \mathbf{I}) \tag{47}$$

$$\overset{perm}{=} (\mathbf{K} \rightrightarrows \mathbf{G}) \rightrightarrows \mathbf{I} \tag{48}$$

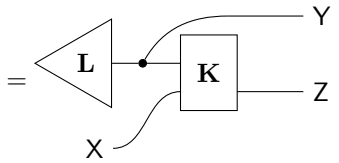
Which implies $\mathbf{J}$ satisfies Equation 23. □

2.6 Submodels

Note that at this point, we have a theory of probability that can handle Equation 1. In particular, $P^{\mathsf{Y}|\mathsf{XZ}}$ must be represented by a labeled Markov kernel $\mathbf{K} : (\mathsf{X}, \mathsf{Z}) \to \Delta(\mathsf{Y})$ and $\mathbb{P}^{\mathsf{Z}}$ by a labeled Markov kernel $\mathbf{L} \in \Delta(\mathsf{Z})$. Then we can define a labeled Markov kernel $\mathbf{M} : \mathsf{X} \to \Delta(\mathsf{X}, \mathsf{Z})$ representing $x \mapsto \mathbb{P}_x^{\mathsf{YZ}}(y, z)$ by

$$\mathbf{M} := \mathbf{L} \rightrightarrows \mathbf{K} \tag{49}$$

$$= \tag{50}$$



We can see that Equation 50 is almost identical to Equation 12 except with set labels instead of sets annotating the wires. This minor change, along with Axiom 1, deals with the problem of identical sets previously mentioned.

However, we do not yet have a notion of *marginal probability* or *conditional probability*. In the paragraph above, the terms $P^{\mathsf{Y}|\mathsf{XZ}}$ and $\mathbb{P}^{\mathsf{Z}}$ are external to the theory developed so far. However, we do know that they are related in a particular way; namely, they are respectively a contitional distribution and a marginal distribution derived from the same probability space with measure P.

We will use the term *submodels* to refer to marginals and conditionals of a higher level model, like those mentioned in the previous paragraph.

Definition 2.5 (Marginal)**.** For any label X, define the marginalising kernel $\overset{*}{\restriction}_{\mathsf{X}} : \mathsf{X} \to *$, which is necessarily unique. Given $\mathbf{K} : \mathsf{W} \to \Delta(\mathsf{Y})$ and $\mathbf{L} : \mathsf{W} \to \Delta(\mathsf{Z})$, L is a *marginal* of $\mathbf{K}$ if, for some X in Y,

$$\mathbf{K} \rightrightarrows \overset{*}{\restriction}_{\mathsf{X}} = \mathbf{L} \tag{51}$$

Because $*$ is in every sequence of labels, $\mathbf{K}$ is always a marginal of itself.

Definition 2.6 (Submodel)**.** Given $\mathbf{K} : \mathsf{X} \to \Delta(\mathsf{Y})$ and $\mathbf{L} : \mathsf{W} \to \Delta(\mathsf{Z})$, $\mathbf{L}$ is a submodel of $\mathbf{K}$ if there are marginals $\mathbf{K}'$, $\mathbf{K}''$ of $\mathbf{K}$ such that

$$\mathbf{K}' \rightrightarrows \mathbf{L} = \mathbf{K}'' \tag{52}$$

As the trivial map $\overset{*}{\restriction} : \mathsf{X} \to \Delta(\{*\})$ is a marginal of $\mathbf{K}$, and $\overset{*}{\restriction} \rightrightarrows \mathbf{K}' = \mathbf{K}'$ for any marginal of $\mathbf{K}$, every marginal of $\mathbf{K}$ is also a submodel of $\mathbf{K}$.

With the definition of submodels in hand, we can introduce a notation more familiar to people experienced with probability theory. If $\mathbf{L} : \mathsf{X} \to \Delta(\mathsf{Y})$ is a submodel of $\mathbf{K}$, we may write $\mathbf{L} \equiv \mathbb{K}^{\mathsf{Y}|\mathsf{X};\mathbf{L}}$ and $\mathbf{L}_x^y \equiv \mathbb{K}^{\mathsf{Y}|\mathsf{X};\mathbf{L}}(y|x)$. Note that the same kernel might be a submodel of many other kernels. This notation isn't *entirely* familiar, as we retain a reference to the original kernel $\mathbf{L}$. In general, a kernel $\mathbf{K}$ has many submodels with the same signature, and this non-uniqueness is more problematic for causal models than for standard probabilistic models, as we will see in Section 4.

Lemma 2.7 (Submodel existence)**.** *In FinStoch, for any* $\mathbf{K} : \mathsf{W} \to \Delta(\mathsf{X}, \mathsf{Y})$ *there exists a submodel* $\mathbb{K}^{\mathsf{Y}|\mathsf{XW};\mathbf{L}}$*.*

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

$$\mathbf{L}^y_{xw} = \frac{\mathbf{K}^{xy}_w}{\sum_{x \in X} \mathbf{K}^{xy}_w} \qquad \forall w, y : \text{ the denominator is positive} \tag{53}$$

Then define $\mathbf{K}^{\mathsf{X}|\mathsf{W};\mathbf{M}} := \mathbf{K} \rightrightarrows {}^{*}\!\!\uparrow_{\mathsf{Y}}$, which is a marginal of $\mathbf{K}$. THen

$$\begin{aligned}
(\mathbf{M} \rightrightarrows \mathbf{L})^{xy}_w &= \mathbf{M}^x_w \mathbf{L}^y_{xw} && (54)\\
&= \sum_{x \in X} \mathbf{K}^{xy}_w \frac{\mathbf{K}^{xy}_w}{\sum_{x \in X} \mathbf{K}^{xy}_w} && \text{if } \mathbf{K}^{xy}_w > 0 \quad (55)\\
&= \mathbf{K}^{xy}_w && \text{if } \mathbf{K}^{xy}_w > 0 \quad (56)\\
&= 0 && \text{otherwise} \quad (57)\\
&= \mathbf{K}^{xy}_w && \text{otherwise} \quad (58)
\end{aligned}$$

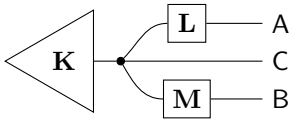
□

We did not use string diagrams in this proof, so this result does not necessarily apply to other Markov kernel categories. This is the technical reason why we choose to work with FinStoch: the existence of submodels presents a challenge in more general settings that we haven't fully resolved and the progress we have is beyond the scope of this paper.

2.7 Conditional independence

We define conditional independence in the following manner:

For a *probability distribution* $\mathbf{P} : \{*\} \to \Delta(\mathsf{Y})$ and some $\mathsf{A}, \mathsf{B}, \mathsf{C} \in \mathsf{Y}$, we say A is independent of B given C, written $\mathsf{A} \perp\!\!\!\perp_{\mathbf{P}} \mathsf{B}|\mathsf{C}$, if there are submodels $\mathbb{P}^{\mathsf{ABC};\mathbf{J}}$, $\mathbb{P}^{\mathsf{C};\mathbf{K}}$, $\mathbb{P}^{\mathsf{A}|\mathsf{C};\mathbf{L}}$, $\mathbb{P}^{\mathsf{B}|\mathsf{C};\mathbf{M}}$ such that

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

(59)

For a *kernel* $\mathbf{N} : \mathsf{X} \to \Delta(\mathsf{Y})$ and some $\mathsf{A}, \mathsf{B}, \mathsf{C} \in (\mathsf{X}, \mathsf{Y})$, we say A is independent of B given C, written $\mathsf{A} \perp\!\!\!\perp_{\mathbf{N}} \mathsf{B}|\mathsf{C}$, if there is some $\mathbf{O} : \{*\} \to \Delta(\mathsf{X})$ such that $O^x > 0$ for all $x \in X$ and $\mathsf{A} \perp\!\!\!\perp_{\mathbf{O} \rightrightarrows \mathbf{N}} \mathsf{B}|\mathsf{C}$.

This definition is inappliccable in the case where sets may be uncountably infinite, as no such $\mathbf{O}$ can exist in this case. There may well be definitions of conditional independence that generalise better, and we refer to the discussions in Fritz (2020) and Constantinou and Dawid (2017) for some discussion of alternative definitions. One advantage of this definition is that it matches the version given by Cho and Jacobs (2019) which they showed coincides with the

standard notion of conditional independence and so we don't have to show this in our particular case.

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

$$\mathbf{K} = \quad \mathsf{X} \text{ ——} \boxed{\mathbf{L}} \text{—— } \mathsf{Y}, \quad \mathsf{W} \text{ ——} * \tag{60}$$

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

3 See-do models

Modular probability is useful when we want to combine different Markov kernels in such a way that "variables" refer to something consistent even though they don't necessarily have a unique distribution. The first example we will present is using modular probability to model decision problems.

Suppose we will be given 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 $\mathbf{S}_\alpha : X \to \Delta(D)$. We are intersted in forecasting some consequences that take values in some set Y, and comparing the forecasts for different strategy choices so as to choose a best strategy.

How can we model this? One way to proceed is as follows: Define a model context $\mathcal{M}$ to which we add the conditional probabilities mentioned hereafter. For each strategy $\mathbb{S}_\alpha[\mathsf{D}|\mathsf{X}]$, our forecast will be represented by some joint probability in $\mathbb{P}_\alpha[\mathsf{XDY}|\mathsf{H}]$ where H is associated with a set of hypotheses H representing different choices that we think might be reasonable to make that may lead to different forecasts. Because observations come before we execute our strategy, we assume that $\mathbb{P}_\alpha[\mathsf{X}|\mathsf{H}] = P_\beta[\mathsf{X}|\mathsf{H}]$ for all α, β. Our chosen strategy is the probability of D given X: $\mathbb{P}_\alpha[\mathsf{D}|\mathsf{X}] \overset{krn}{=} \mathbb{S}_\alpha[\mathsf{D}|\mathsf{X}]$. Finally, our forecast of Y is the same for all strategies holding the observations, the decision and the hypothesis fixed: $\mathbb{P}_\alpha[\mathsf{Y}|\mathsf{HD}] = P_\beta[\mathsf{Y}|\mathsf{HD}]$ for all α, β.

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

$$\mathbb{P}_\alpha[\mathsf{XDY}|\mathsf{H}] \overset{krn}{=} \mathbb{T}[\mathsf{X}|\mathsf{H}] \rightrightarrows \mathbb{S}_\alpha[\mathsf{D}|\mathsf{X}] \rightrightarrows \mathbb{T}[\mathsf{Y}|\mathsf{XHD}] \tag{61}$$

The proof is given in Appendix 6. Note that $\mathbb{T}[\mathsf{X}|\mathsf{H}]$ exists by virtue of the fact $\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$. While this independence is what enables Equation 61, in general $\mathsf{X} \not\perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{D}|\mathsf{H}$, so $\mathbb{T}$ cannot be a disintegration of $\mathbb{P}_\alpha$. Modular probability allows us to specify $\mathbb{T}$, which we call a *see-do model*, as a partial forecast to be completed with a strategy $\mathbb{S}_\alpha$ 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 $\mathbb{P}_\alpha$, $\mathbb{T}$, 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 Θ. We can observe that $\{\mathbb{T}[\mathsf{X}|\mathsf{H}]_h\}_{h\in H}$ 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 $\mathbb{S}_\alpha : X \to \Delta(D)$ is evaluated according to the risk functional $R(\theta, \mathbb{S}_\alpha) = \sum_{x\in X}\sum_{d\in D} \mathbb{P}^x_\theta(S_\alpha)^d_x l(h, d)$.

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

$$\mathbb{E}_{\mathbb{P}_\alpha[\mathsf{XDY}|\mathsf{H}]}[\mathsf{Y}] = \sum_{x\in X}\sum_{d\in D}\sum_{y\in Y} y\,(\mathbb{T}[\mathsf{X}|\mathsf{H}] \rightrightarrows \mathbb{S}_\alpha[\mathsf{D}|\mathsf{X}] \rightrightarrows \mathbb{T}[\mathsf{Y}|\mathsf{XHD}])^{xdy}_h \tag{62}$$

$$= \sum_{x\in X}\sum_{d\in D}\sum_{y\in Y} \mathbb{T}[\mathsf{X}|\mathsf{H}]^x_h \mathbb{S}_\alpha[\mathsf{D}|\mathsf{X}]^d_x \mathbb{T}[\mathsf{Y}|\mathsf{HD}]^y_{h,d} \tag{63}$$

$$= \sum_{x\in X}\sum_{d\in D} \mathbb{T}[\mathsf{X}|\mathsf{H}]^x_h (S_\alpha)^d_x l(h, d) \tag{64}$$

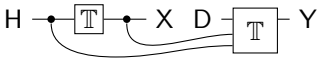
$$= R(h, \mathbb{S}_\alpha) \tag{65}$$

That is, if we are given a see-do model where we interpret $\mathbb{T}[\mathsf{X}|\mathsf{H}]$ as a statistical experiment and Y as a reversed utility, the expectation of the utility under the strategy forecast given in equation 61 is the risk of that strategy under hypothesis h.

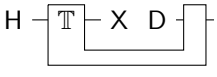
3.2 Combs

The see-do model $\mathbb{T}[\mathsf{XY}|\mathsf{HD}]$ 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 61), 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:

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

(66)

$=$ 

(67)

We can take any strategy $\mathbb{S}_\alpha[\mathsf{D}|\mathsf{X}]$ and drop it into the "hole" in 67 (as described in Equation 61) to get a forecast of the outcome of that strategy.

4 Causal Bayesian Networks

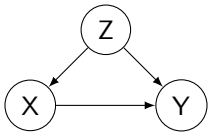
A Causal Bayesian Network posits a set of observational probability distributions, for example $\{\mathbb{P}_h(X,Y)\}_H$, and a set of interventional distributions, for example $\{\mathbb{P}_h(X,Y|do(X=x))\}_{x\in X,h\in H}$. Here we use notation similar to typical notation used for Causal Bayesian Networks 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 X,Y and call the latter X',Y'. Let $D=\{do(\mathsf{X}=x)\}_{x\in X}$. Then a Causal Bayesian Network can be considered a see-do model $\mathbb{T}[\mathsf{XYX'Y'|HD}]$ by identifying $\mathbb{T}[\mathsf{XY|H}]_h := \mathbb{P}_h(X,Y)$ and $\mathbb{T}[\mathsf{X'Y'|HD}]_{h,do(\mathsf{X}=x)} := P_h(X,Y|do(X=x))$.

We need to rename the consequence variables because otherwise we would have $\mathbb{T}[\mathsf{XYXY|HD}]$ and the two X's and the two 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 $\mathbb{T}[\mathsf{XYZ|H}]$ be the observational "see" model and $\mathbb{T}[\mathsf{X'Y'Z'|HD}]$ 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 $\mathbb{T}[\mathsf{X'Y'Z'|HD}]^{x'yz}_{hx} \overset{krn}{=} \mathbb{T}[\mathsf{Z|H}]^{z}_{h}\delta[x]^{x'}\mathbb{T}[\mathsf{Y|XZH}]^{y}_{hx'z}$.

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

$$\mathbb{T}[\mathsf{XYZ|H}] = \mathbb{U}[\mathsf{Z|H}] \rightrightarrows \mathbb{T}[\mathsf{X|H}] \rightrightarrows \mathbb{U}[\mathsf{Y|XHZ}] \tag{68}$$

$$\mathbb{T}[\mathsf{X'Y'Z'|HD}] \overset{krn}{=} \mathbb{U}[\mathsf{Z|H}] \rightrightarrows \mathbb{D}[\mathsf{X|D}] \rightrightarrows \mathbb{U}[\mathsf{Y|XHZ}] \tag{69}$$

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

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

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

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

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

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

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

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

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

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

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

5.1 Potential outcomes in see-do models

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

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

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

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

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

As an aside that we could define stochastic potential outcomes if we allow the variables $\mathsf{Y}(x)$ to take values in $\Delta(Y)$ rather than in Y, and then require $\mathbf{K}[\mathsf{Y}|\mathsf{X}\mathsf{Y}(X)]_{ij_1j_2\dots j_{|X|}} = j_i$ (where j_i is an element of $\Delta(Y)$). This is more

complex to work with and rarely seen in practice, but it is worth noting that Definition 5.1 can be generalised to cover models where $\mathsf{Y}(x)$ describes the value Y would take if X were x *with uncertainty*.

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

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

Definition 5.2 (Parallel potential outcomes)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, a collection of variables $\{\mathsf{Y}_i, \mathsf{Y}(W), \mathsf{W}_i\}$, $i \in [n]$, where Y_i and $\mathsf{Y}(W)$ are random variables and W_i could be either a state or random variables is a *parallel potential outcome submodel* if $\mathbf{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]$ exists and $\mathbf{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]_{kj_1j_2\ldots j_{|W|}} = \delta[j_k]$.

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

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

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

5.2 Parallel potential outcomes representation theorem

Exchangeble sequences of random variables are sequences whose joint distribution is unchanged by permutation. Independent and identically distributed random variables are one example: if X_1 is the result of the first flip of a coin

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

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

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

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

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

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

Definition 5.3 (Exchangeability)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ with Y_i random variables is *exchangeable* conditional on Z if, defining $\mathsf{Y}_{[n]} = (\mathsf{Y}_i)_{i\in[n]}$ and $\mathsf{D}_{[n]} = (\mathsf{D}_i)_{i\in[n]}$, $\mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$ exists and for any bijection $\pi : [n] \to [n]$ $\mathbf{K}[\mathsf{Y}_{\pi([n])}|\mathsf{D}_{\pi([n])}\mathsf{Z}] = \mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$.

Definition 5.4 (Extension)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, $(\mathbf{K}', E', F')$ is an *extension* of $(\mathbf{K}, E, F)$ if there is some random variable X and some state variable U such that $\mathbf{K}'[\mathsf{X}|\mathsf{U}]$ exists and $\mathbf{K}'[\mathsf{X}|\mathsf{U}] = \mathbf{K}$.

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

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

Definition 5.5 (Extendably exchangeable)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ and a state variable Z with Y_i random variables is *extendably exchangeable* if there exists an extension of $\mathbf{K}$ with respect to which $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in\mathbb{N}}$ is exchangeable conditional on Z.

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

Deterministically reproducible sequences have the property that repeating the same decision gets the same response with probability 1. This could be a model of an experiment that exhibits no variation in results (e.g. every time I put green paint on the page, the page appears green), or an assumption about collections of "what-ifs" (e.g. if I went for a walk an hour ago, just as I actually did, then I definitely would have stubbed my toe, just like I actually did). Incidentally, many consider that this assumption is false concering what-if questions about things that exhibit quantum behaviour.

Definition 5.6 (Deterministically reproducible)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ with Y_i random variables is *deterministically reproducible* conditional on Z if $n \geq 2$, $\mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$ exists and $\mathbf{K}[\mathsf{Y}_{\{i,j\}}|\mathsf{D}_{\{i,j\}}\mathsf{Z}]^{lm}_{kk} = [\![l = m]\!]\mathbf{K}[\mathsf{Y}_i|\mathsf{D}_i\mathsf{Z}]^l_k$ for all i, j, k, l, m.

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

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

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

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

$$\mathbf{K}[\mathsf{X}_n|\mathsf{D}_n\mathsf{Z}]^x_{dz} = \sum_{y \in Y^{|D|}} \mathbf{K}[\mathsf{X}_{[n]}|\mathsf{D}_n\mathsf{Z}\mathsf{Y}(D)]^x_{dzy}\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_{[n]}\mathsf{Z}]^y_{dz} \tag{70}$$

$$= \sum_{y \in Y^{|D|}} \prod_{i \in [n]} \delta[y_{d_i}]^{x_i}\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_n\mathsf{Z}]^y_{dz} \tag{71}$$

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

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

$$\mathbf{K}[\mathsf{X}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]^x_{dz} = \sum_{y \in Y^{|D|}} \prod_{i \in Q_d} \delta[y_{d_i}]^{x_i}\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_{[n]}\mathsf{Z}]^y_{dz} \tag{72}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}(R_d)\mathsf{Y}(R_d^C)|\mathsf{D}_{Q_d}\mathsf{D}_{Q_d^C}\mathsf{Z}]^{x_{Q_d}y'}_{d_{Q_d}d^C_{Q_d}z} \tag{73}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}_{R_d}\mathsf{Y}_{R_d^C}|\mathsf{D}_{Q_d}\mathsf{D}_{Q_d^C}\mathsf{Z}]^{x_{Q_d}y'}_{dz} \tag{74}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]^{x_{Q_d}y'}_{dz} \qquad \text{(using exchangeability)} \tag{75}$$

Note that

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

6 Appendix:see-do model representation

Modularise the treatment of probability

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

set of hypotheses H, and under each hypothesis $h \in H$ we model the result of choosing some strategy $\mathbf{S}$ as a joint probability over observations, decisions and consequences $\mathbb{P}_{h,\mathbf{S}} \in \Delta(X \times D \times Y)$.

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

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

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

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

$$\mathbb{P}[\mathsf{XDY}]^{ijk} = \mathbb{P}[\mathsf{X}]^i \mathbb{P}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{P}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \tag{76}$$

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

$$\mathbf{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^l_{kj} \tag{77}$$

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

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

We then have

$$\begin{aligned}
\mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj} &= \mathbf{T}\underline{\mathsf{X}'}^k_{hj} && (78)\\
&= \sum_l \mathbf{T}^{kl}_{hj} && (79)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^k && (80)\\
&= \mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj'} && (81)
\end{aligned}$$

Thus $\mathsf{X}' \perp\!\!\!\perp_{\mathbf{T}} \mathsf{D}'|\mathsf{H}'$ and so $\mathbf{T}[\mathsf{X}'|\mathsf{H}']$ exists (section 2.7) and $(\mathbf{T}, \mathsf{H}', \mathsf{D}', \mathsf{X}', \mathsf{Y}')$ is a see-do model.

Applying Equation 76 to $\mathbb{P}_{h,\mathbf{S}}$:

$$\begin{aligned}
\mathbb{P}_{h,\mathbf{S}}[\mathsf{XDY}]^{ijk} &= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^{i}\mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]^{j}_{i}\mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^{k}_{ij} && (82)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^{i}\mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^{k}_{ij} && (83)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]^{j}_{i}\mathbf{T}[\mathsf{X'Y'}|\mathsf{H'D'}]^{ik}_{hj} && (84)\\
&= \mathbf{S}^{j}_{i}\mathbf{T}[\mathsf{X'Y'}|\mathsf{H'D'}]^{ik}_{hj} && (85)\\
&= \mathbf{S}^{j}_{i}\mathbf{T}[\mathsf{X'}|\mathsf{H'D'}]^{i}_{hj}\mathbf{T}[\mathsf{Y'}|\mathsf{X'H'D'}]^{k}_{ihj} && (86)\\
&= \mathbf{T}[\mathsf{X'}|\mathsf{H'}]^{i}_{h}\mathbf{S}^{j}_{i}\mathbf{T}[\mathsf{Y'}|\mathsf{X'H'D'}]^{k}_{ihj} && (87)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^{i} &= \mathbb{P}_{h,\mathbf{Q}}[\mathsf{X}]^{i} && (88)\\
\mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^{k}_{ij} &= \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}]^{k}_{ij} \text{ for some version of } \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}] && (89)
\end{aligned}$$

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

$$\mathbf{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbf{Q}}[\mathsf{X}]^{k}\mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}]^{l}_{kj} \tag{90}$$

Then by substitution of $\mathbf{Q}$ for $\mathbf{S}$ in Equation 82 and working through the same steps

$$\mathbb{P}_{h,\mathbf{S}}[\mathsf{XDY}]^{ijk} = \mathbf{T}[\mathsf{X'}|\mathsf{H'}]^{i}_{h}\mathbf{Q}^{j}_{i}\mathbf{T}[\mathsf{Y'}|\mathsf{X'H'D'}]^{k}_{ihj} \tag{91}$$

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 20 in other Markov kernel categories.

Define

$$\mathsf{I}_{K\cdot\cdot} := \mathsf{I}_K \setminus \mathsf{I}_L \setminus \mathsf{I}_J \tag{92}$$
$$\mathsf{I}_{KL\cdot} := \mathsf{I}_K \cap \mathsf{I}_L \setminus \mathsf{I}_J \tag{93}$$
$$\mathsf{I}_{K\cdot J} := \mathsf{I}_K \cap \mathsf{I}_J \setminus \mathsf{I}_L \tag{94}$$
$$\mathsf{I}_{KLJ} := \mathsf{I}_K \cap \mathsf{I}_L \cap \mathsf{I}_J \tag{95}$$
$$\mathsf{I}_{\cdot L\cdot} := \mathsf{I}_L \setminus \mathsf{I}_K \setminus \mathsf{I}_J \tag{96}$$
$$\mathsf{I}_{\cdot LJ} := \mathsf{I}_L \cap \mathsf{I}_J \setminus \mathsf{I}_K \tag{97}$$
$$\mathsf{I}_{\cdot\cdot J} := \mathsf{I}_J \setminus \mathsf{I}_K \setminus \mathsf{I}_L \tag{98}$$
$$\mathsf{O}_{K\cdot\cdot} := \mathsf{O}_K \setminus \mathsf{I}_N \setminus \mathsf{I}_J \tag{99}$$
$$\mathsf{O}_{KL\cdot} := \mathsf{O}_K \cap \mathsf{I}_L \setminus \mathsf{I}_J \tag{100}$$
$$\mathsf{O}_{K\cdot J} := \mathsf{O}_K \cap \mathsf{I}_J \setminus \mathsf{I}_L \tag{101}$$
$$\mathsf{O}_{KLJ} := \mathsf{O}_K \cap \mathsf{I}_L \cap \mathsf{I}_J \tag{102}$$
$$\mathsf{O}_{L\cdot} := \mathsf{O}_L \setminus \mathsf{I}_J \tag{103}$$
$$\mathsf{O}_{LJ} := \mathsf{O}_L \cap \mathsf{I}_J \tag{104}$$

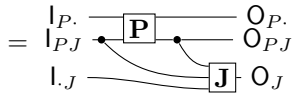
Also define

$$(\mathbf{P}, \mathsf{I}_P, \mathsf{O}_P) := \mathbf{K} \rightrightarrows \mathbf{L} \tag{105}$$
$$(\mathbf{Q}, \mathsf{I}_Q, \mathsf{O}_Q) := \mathbf{L} \rightrightarrows \mathbf{J} \tag{106}$$

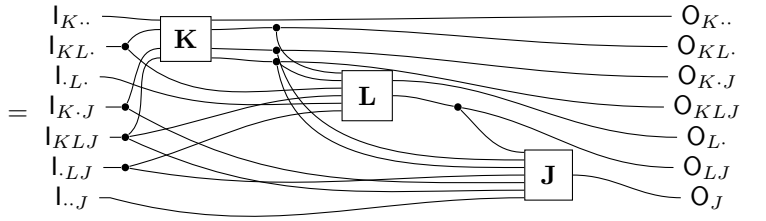
Then

$$(\mathbf{K} \rightrightarrows \mathbf{L}) \rightrightarrows \mathbf{J} = \mathbf{P} \rightrightarrows \mathbf{J} \tag{107}$$

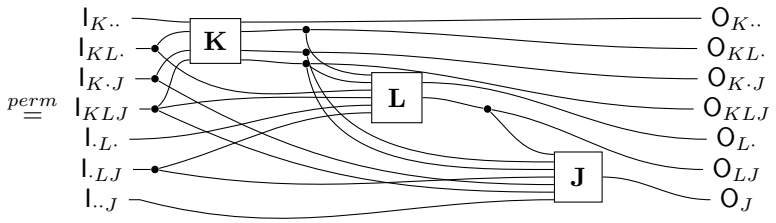
$$= \tag{108}$$



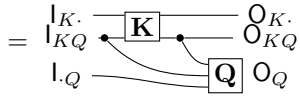
$$= \tag{109}$$



$$\overset{perm}{=} \tag{110}$$



$$= \tag{111}$$

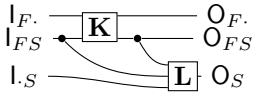


$$= \mathbf{K} \rightrightarrows (\mathbf{L} \rightrightarrows \mathbf{J}) \tag{112}$$

8 Appendix: String Diagram Examples

Recall the definition of *connection*:

$$\mathbf{K} \rightrightarrows \mathbf{L} := \tag{113}$$



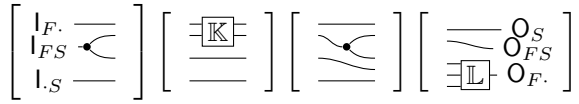
$$:= \mathbf{J} \tag{114}$$

$$\mathbf{J}^{zxw}_{yqr} = \mathbf{K}^{zx}_{yq}\mathbf{L}^{w}_{xqr} \tag{115}$$

Equation 113 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:

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



(116)

(117)

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.

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`.

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.

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: