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

David Johnston

November 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.

---

[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.

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

- We need to talk about variables

- We use compatibility + string diagrams

- We consider causation in terms of "proxy control"

2   Variables and Probability Models

2.1   Probability distributions, Markov kernels and string diagrams

We make extensive use of probability theory, and the following is a brief introduction to the string diagram notation we use for probabilistic reasoning. This notation comes from the study of Markov categories. Markov categories are abstract categories that represent models of the flow of information. We can form Markov categories from collections of sets – for example, discrete sets or standard measurable sets – along with the Markov kernel product as the composition operation. Markov categories come equipped with a graphical language of *string diagrams*, and a coherence theorem which states that calid proofs using string diagrams correspond to valid theorems in *any* Markov category (Selinger, 2010). More comprehensive introductions to Markov categories can be found in Fritz (2020); Cho and Jacobs (2019). Thus, while we limit ourselves to discrete sets in this paper, any derivation that uses only string diagrams is more broadly appliccable.

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

1. $y \mapsto \mathbb{Q}^{\mathsf{X}|\mathsf{Y}}(A|y)$ is $\mathcal{B}$-measurable for all $A \in \mathcal{X}$

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

In the context of discrete sets, a probability distribution can be defined as a vector, and a Markov kernel a matrix.

Definition 2.1 (Probability distribution (discrete sets))**.** A probability distribution $\mathbb{P}$ on a discrete set X is a vector $(\mathbb{P}(x))_{x \in X} \in [0, 1]^{|X|}$ such that $\sum_{x \in X} \mathbb{P}(x) = 1$. For $A \subset X$, define $\mathbb{P}(A) = \sum_{x \in A} P(x)$.

Definition 2.2 (Markov kernel (discrete sets))**.** A Markov kernel $\mathbb{K} : X \rightarrow Y$ is a matrix $(\mathbb{K}(y|x))_{x \in X, y \in Y} \in [0, 1]^{|X||Y|}$ such that $\sum_{y \in Y} \mathbb{K}(y|x) = 1$ for all $x \in X$. For $B \subset Y$ define $\mathbb{K}(B|x) = \sum_{y \in B} \mathbb{K}(y|x)$.

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

$$\mathbb{K} := \quad -\boxed{\mathbb{K}}- \tag{1}$$

$$\mathbb{P} := \quad \triangleleft\!\boxed{\mathbb{P}}- \tag{2}$$

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

$$\mathbb{L}\mathbb{M} := \quad X \boxed{\mathbb{K}}\!-\!\boxed{\mathbb{M}} \, Z \tag{3}$$

The Cartesian product $X \times Y := \{(x, y) | x \in X, y \in Y\}$. Given kernels $\mathbb{K} : W \rightarrow Y$ and $\mathbb{L} : X \rightarrow Z$, the tensor product $\mathbb{K} \otimes \mathbb{L} : W \times X \rightarrow Y \times Z$ given by $(\mathbb{K} \otimes \mathbb{L})(y, z | w, x) := K(y|w)L(z|x)$. The tensor product is graphically represeted by drawing kernels in parallel:

$$\mathbb{K} \otimes \mathbb{L} := \quad \begin{matrix} W \, \boxed{\mathbb{K}} \, Y \\ X \, \boxed{\mathbb{L}} \, Z \end{matrix} \tag{4}$$

We read diagrams from left to right (this is somewhat different to Fritz (2020); Cho and Jacobs (2019); Fong (2013) but in line with Selinger (2010)),

and any diagram describes a set of nested products and tensor products of Markov kernels. There are a collection of special Markov kernels for which we can replace the generic "box" of a Markov kernel with a diagrammatic elements that are visually suggestive of what these kernels accomplish.

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

$$\mathrm{id}_X := \quad X \text{---} X \tag{5}$$

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

$$\mathrm{del}_X := \quad \text{---} * \, X \tag{6}$$

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

$$\mathrm{copy}_X := \quad X \prec{\begin{matrix} X \\ X \end{matrix}} \tag{7}$$

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

$$\mathrm{swap}_X := \quad \times \tag{8}$$

Because we anticipate that the graphical notation will be unfamiliar, we will include some examples in the next section.

2.1.1 Examples

When translating string diagram notation to integral notation, a number of identities can speed up the process.

For arbitrary $\mathbb{K} : X \times Y \to Z$, $\mathbb{L} : W \to Y$

$$[(\mathrm{id}_X \otimes \mathbb{L})\mathbb{K}](A|x, w) = \int_Y \int_X \mathbb{K}(z|x', y')\mathbb{L}(dy'|w)\mathrm{id}_X(dx'|x) \tag{9}$$

$$= \int_Y \mathbb{K}(z|x, y')\mathbb{L}(dy'|w) \tag{10}$$

That is, an identity map passes its input to the next kernel in the product.

For arbitrary $\mathbb{K} : X \times Y \times Y \to Z$ (where we apply the above shorthand in the first line):

$$[(\mathrm{id}_X \otimes \mathrm{copy}_Y)\mathbb{K}](A|x,y) = \int_Y \int_Y \mathbb{K}(A|x,y',y'')\mathrm{copy}_Y(dy' \times dy''|y) \tag{11}$$
$$= \mathbb{K}(A|x,y,y) \tag{12}$$

That is, the copy map passes along two copies of its input to the next kernel in the product.

For a collection of kernels $\mathbb{K}^n : Y^n \to Z$, $n \in [n]$, define $(y)^n = (y|i \in [n])$ and:

$$\mathrm{copy}_Y^n := \begin{cases} \mathrm{copy}_Y^{n-1}(\mathrm{id}_{Y^{n-2}} \otimes \mathrm{copy}_Y) & n > 2 \\ \mathrm{copy}_Y & n = 2 \end{cases} \tag{13}$$
$$(\mathrm{copy}_Y^2 \mathbb{K}^2)(z|y) = \mathbb{K}^2(z|y,y) \tag{14}$$
$$\tag{15}$$

Suppose for induction

$$(\mathrm{copy}_Y^{n-1}\mathbb{K}^{n-1})(z|y) = \mathbb{K}^{n-1}(z|(y)^{n-1}) \tag{16}$$

then

$$(\mathrm{copy}_Y^n \mathbb{K}^n)(z|y) = (\mathrm{copy}_Y^{n-1}(\mathrm{id}_{Y^{n-2}} \otimes \mathrm{copy}_Y)\mathbb{K}^n)(z|y) \tag{17}$$
$$= \sum_{y' \in Y^{n-1}} (\mathrm{id}_{Y^{n-2}} \otimes \mathrm{copy}_Y)(\mathbf{y}'|(y)^{n-1})\mathbb{K}^n(z|\mathbf{y}') \tag{18}$$
$$= \mathbb{K}^n(z|(y)^n) \tag{19}$$
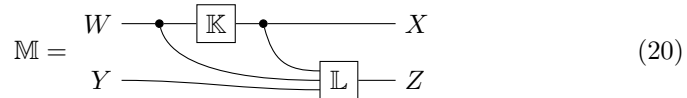
That is, we can define the n-fold copy map that passes along n copies of its input to the next kernel in the product.

2.1.2 Example: comb insertion

The following examples illustrate 2-combs and the insertion operation, both of which we will define later. As an example in translating diagrams, we show how the diagrams for a 2-comb and 2-comb with an inserted Markov kernel can be translated to integral notation.

Consider the Markov kernels $\mathbb{K} : W \rightarrow X$, $\mathbb{L} : X \times W \times Y \rightarrow Z$ and the 2-comb $\mathbb{M} : W \times Y \to X \times Z$ defined as

$$\tag{20}$$

6

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

$$\mathbb{M}(A \times B|w,y) = [(\text{copy}_W \otimes \text{id}_Y)(\mathbb{K} \otimes \text{id}_{W \times Y}) \tag{21}$$
$$(\text{copy}_X \otimes \text{id}_{W \times Y})(\text{id}_X \otimes \mathbb{L})](A \times B|w,y) \tag{22}$$
$$= [(\mathbb{K} \otimes \text{id}_{W \times Y})(\text{copy}_X \otimes \text{id}_{W \times Y}) \tag{23}$$
$$(\text{id}_X \otimes \mathbb{L})](A \times B|w,w,y) \tag{24}$$
$$= \int_X (\text{id}_X \otimes \mathbb{L})(A \times B|x',w,y)\mathbb{K}(dx'|w)](y,z|y',x) \tag{25}$$
$$= \int_X \text{id}_X(A|x')\mathbb{L}(B|x',w,y)\mathbb{K}(dx'|w) \tag{26}$$
$$= \int_A \mathbb{L}(B|x',w,y)\mathbb{K}(dx'|w) \tag{27}$$

If we are given additionally $\mathbb{J} : X \times W \rightarrow Y$, we can define a new Markov kernel $\mathbb{N} : W \rightarrow Z$ given by "inserting" $\mathbb{J}$ into $\mathbb{M}$:

$$\tag{28}$$
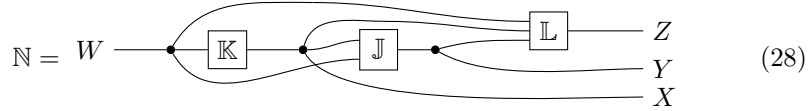
We can translate Equation 28 to

$$\mathbb{N}(A \times B \times C|w) = [\text{copy}_W(\mathbb{K}\text{copy}_Y^3 \otimes \text{id}_W) \tag{29}$$
$$(\text{id}_Y \otimes \mathbb{J} \otimes \text{id}_Y)(\text{id}_Y \otimes \text{copy}_X \otimes \text{id}_Y) \tag{30}$$
$$(\mathbb{L} \otimes \text{id}_X \otimes \text{id}_Y)](A \times B \times C|w) \tag{31}$$
$$= [(\mathbb{K}\text{copy}_Y^3 \otimes \text{id}_W)(\text{id}_Y \otimes \mathbb{J} \otimes \text{id}_Y) \tag{32}$$
$$(\text{id}_Y \otimes \text{copy}_X \otimes \text{id}_Y) \tag{33}$$
$$(\mathbb{L} \otimes \text{id}_X \otimes \text{id}_Y)](A \times B \times C|w,w) \tag{34}$$
$$= \int_X \int_Y \mathbb{L}(C|x',w,y')\text{id}_X(A|x')\text{id}_Y(B|y')\mathbb{J}(dy'|x',w)\mathbb{K}(dx'|w) \tag{35}$$
$$= \int_A \int_B \mathbb{L}(C|x',w,y')\mathbb{J}(dy'|x',w)\mathbb{K}(dx'|w) \tag{36}$$

2.2 Semantics of observed and unobserved variables

We are interested in constructing *probabilistic models* which explain some part of the world. In a model, variables play the role of "pointing to the parts of

the world the model is explaining". Both observed an unobserved variables play important roles in causal modelling and we think it is worth clarifying what variables of either type refer to. Our approach is a standard one: a probabilistic model is associated with an experiment or measurement procedure that yields values in a well-defined set. Observable variables are obtained by applying well-defined functions to the result of this total measurement. We use a richer sample space that includes "unobserved variables" that are formally treated the same way as observed variables, but aren't associated with any real-world counterparts.

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

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

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

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

$$(g \circ f) \circ \mathcal{B} \text{ yields } x \iff B \text{ yields } (g \circ f)^{-1}(x) \tag{37}$$
$$\iff B \text{ yields } f^{-1}(g^{-1}(x)) \tag{38}$$
$$\iff f \circ B \text{ yields } g^{-1}(x) \tag{39}$$
$$\iff g \circ (f \circ B) \text{ yields } x \tag{40}$$

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

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

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

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

$\omega \in \Omega$ is *consistent with observation* iff the result yielded by $\mathcal{S}$ is equal to $\mathsf{S}(\omega)$
$$\tag{41}$$

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

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

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

By a *variable* we will mean an attribute, measurement or inquiry that may take on one of several possible outcomes, or values, from a specified domain. If we have beliefs (i.e., probabilities) attached to the possible values that a variable may attain, we will call that variable a random variable.

This is a minor generalization of the textbook definition, according to which a random variable is a mapping from the sample space (e.g., the set of elementary events) to the real line. In our definition, the mapping is from the sample space to any set of objects called "values," which may or may not be ordered.

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

2.3 Events

To recap, we have a procedure $\mathcal{S}$ yielding values in Ψ that measures everything we are interested in, a sample space Ω and a function S that models $\mathcal{S}$ in the sense of Definition 41. We assume also that Ψ has a σ-algebra $\mathcal{E}$ (this may be the power set of Ψ, as measurement procedures are typically limited to finite precision). Ω is equipped with a σ-algebra $\mathcal{F}$ such that $\sigma(\mathsf{S}) \subset \mathcal{F}$. If a procedure $\mathcal{X} = f \circ \mathcal{S}$ then we define $\mathsf{X} : \Omega \to X$ by $\mathsf{X} := f \circ \mathsf{S}$.

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

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

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

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

2.4 Probabilistic models for causal inference

The sample space $(\Omega, \mathcal{F})$ along with our collection of variables is a "model skeleton" – it tells us what kind of data we might see. The process $\mathcal{S}$ which tells

us which part of the world we're interested in is related to the model Ω and the observable variables by the criterion of *consistency with observation*. The kind of problem we are mainly interested in here is one where we make use of data to help make decisions under uncertainty. Probabilistic models have a long history of being used for this purpose, and our interest here is in constructing probabilistic models that can be attached to our variable "skeleton".

For causal inference, we find we need to generalise the standard approach to constructing probability models on a given sample space $(\Omega, \mathcal{F})$. The key things we need to handle are *gaps* in our model. Hájek (2003) defines *probability gaps* as propositions that do not have a probability assigned to them. Our view of probability gaps is slightly different – in this work, a model with probability gaps as one that is missing some key parts. If we complete such a model with parts of the appropriate type, we get a standard probability model.

Probability gaps are useful in models used for decision making because, when I have a number of different options I could choose, a model can only help select from them if it tells me what is likely to happen for each choice I could make. Thus if we have a variable representing choices, we must have a model that can tolerate a probability gap for this variable.

As an initial example of a probability gap in causal inference, we will consider the example of truncated factorisation. For this example, we will assume that the reader is familiar enough with marginal probabilities, conditional probabilities and causal models to follow along. We will offer more careful definitions of terms later.

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

$$\mathbb{P}^{\mathsf{XYZ}}_x(x', y, z) = \mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}(y|x, z)\mathbb{P}^{\mathsf{Z}}(z)[\![x = x']\!] \tag{42}$$

Equation 42 embodies three assumptions about a model of the operation of "setting X to x". First, such a model must assign probability 1 to the proposition that X yields x. Second, such a model must assign the same marginal probability distribution to Z as the input distribution; $\mathbb{P}^{\mathsf{Z}} = \mathbb{P}^{\mathsf{Z}}_x$. Finally, our model must also assign the same conditional probability to Y given X and Z; $\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}} = \mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}_x$.

Notice that the map $x \mapsto \mathbb{P}^{\mathsf{XYZ}}_x$ itself "looks like" a conditional probability. It maps each $x \in X$ to a probability distribution over $(\mathsf{X}, \mathsf{Y}, \mathsf{Z})$. In fact, a popular alternative notation for $\mathbb{P}^{\mathsf{XYZ}}_x$ this map is $\mathbb{P}^{\mathsf{XYZ}|do(\mathsf{X}=x)}$, which is clearly suggestive of an interpretation as a kind of conditional probability. We will take this interpretation seriously: we will posit some variable U (which may or may not be observable) and a probabilisitic model $\mathbb{Q}^{\mathsf{XYZ}|\mathsf{U}} := x \mapsto \mathbb{P}^{\mathsf{XYZ}}_x$.

We note that the thing called $do(\mathsf{X} = x)$ or that we have called U is often referred to as an *intervention*. Interventions are often things we can choose to

do or not to do. Also, perhaps, we could also consider choosing to do or not do an intervention based on the output of some random process. In this case, we will need a model that can tell us which result we are likely to see for any choice of $\mathbb{Q}^{\mathsf{U}}_\alpha$; the distribution of U is a *probability gap*.

$\mathbb{Q}^{\mathsf{XYZ}|\mathsf{U}}$, as we have defined it so far, is not quite an ideal candidate for for a gappy probability model. Firstly, because conditional probabilities are arbitrary on sets of measure zero with regard to $\mathbb{P}^{\mathsf{XYZ}}$, definition 42 can be satisfied by multiple probability distributions that differ in meaningful ways. Suppose X, Y and Z are binary, $\mathbb{P}^{\mathsf{Z}}(1) = 1$ and $\mathbb{P}(\mathsf{X} \bowtie \mathsf{Z}) = 1$. Then we can consistently choose $\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}(1|0,1) = 1$ or $\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}(1|0,1) = 0$ because $\{0,1\}$ is a measure zero event. However, the first choice gives us $\mathbb{P}^{\mathsf{XYZ}}_0(0,1,1) = 1$ while the second gives us $\mathbb{P}^{\mathsf{XYZ}}_0(0,1,1) = 0$, which are very different opinions regarding "the result of setting X to 1".

Secondly, there may be no probability model at all that satisfies Equation 42. For example, suppose $\mathsf{X} = f \circ \mathsf{Z}$ for some f. Then we must have $\mathbb{P}^{\mathsf{X}}_x(x') = \mathbb{P}^{\mathsf{Z}}_x(f^{-1}(x'))$ for any x. However, we also have $\mathbb{P}^{\mathsf{X}}_x(x') = [\![x = x']\!]$ for all x, x' and $\mathbb{P}^{\mathsf{Z}}_x = \mathbb{P}^{\mathsf{Z}}$ for all x. Thus if X can more than one value, there is at least one choice of x that cannot simultaneously satisfy these requirements.

This might seem like an absurd model, but an analogous causal graph appears in Shahar (2009) where $\mathsf{Z} = (\mathsf{H}, \mathsf{W})$, representing a person's height and weight, and X represents their body mass index, which is to say $\mathsf{X} = \frac{\mathsf{W}}{\mathsf{H}^2}$. Furthermore, this paper uses this model to argue that body mass index cannot have a causal effect. Such an argument cannot be supported by Equation 42 because by that equation there is no probability model corresponding to an intervention on X.

Our first aim is to offer a more careful theory of probability models with gaps in them, addressing these two shortcomings.

2.5 Probability gaps

Probability gap models are probability models that are missing key parts. If we provide a "missing part" or "insert" of the right type then we get a probability model. We can think of probability as functions that map inserts to probability models, or as higher-order functions that map inserts to lower order probability gap models.

Not every function of this type is a probability gap model; the probability model you get in the end has to break into a "fixed part", which stays the same regardless of the insert provided and additionally the insert. For a concrete example: a conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ is a kind of probability gap model, and when we provide an insert $\mathbb{Q}^{\mathsf{X}}_\alpha$ we get a probability model $\mathbb{P}^{\mathsf{XY}}_\alpha$ which depends on the $\mathbb{Q}^{\mathsf{X}}_\alpha$ chosen. In turn, $\mathbb{P}^{\mathsf{XY}}_\alpha$ contains the insert $\mathbb{Q}^{\mathsf{X}}_\alpha$ via marginalisation and the fixed $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ via disintegration.

Probability gap models must be *valid*. A conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ is a Markov kernel $X \rightarrowtail Y$, but not every Markov kernel $X \rightarrowtail Y$ can be extended to a probability model. Those that can extend to a probability model, given a valid insert, are valid probability gap models.

We will make a substantial simplifying assumption: all sets, including the sample space Ω and any set of values a variable takes, are discrete sets. That is, they are at most countably infinite and the σ-algebra of measurable sets is the power set. Because we are working with discrete sets we will by convention call probability measures on set elements: $\mathbb{P}^{\mathsf{X}}(x) := \mathbb{P}^{\mathsf{X}}(\{x\})$.

Definition 2.3 (Probability space)**.** A probability space is a triple $(\mathbb{P}^{\Omega}, \Omega, \mathcal{F})$, where μ is a probability measure on $\mathcal{F}$ which, as stated above, we take to be $\mathcal{P}(\Omega)$. We call $\mathbb{P}^{\Omega}$ the *base measure*, and the superscript Ω is a convention we use to name base measures.

Probability spaces along with random variables can be used to define probability distributions of those random variables. We call such distributions *pushforwards*.

Definition 2.4 (Pushforward)**.** Given a probability space and a random variable $\mathsf{X} : \Omega \to X$, we can define the probability distribution $\mathbb{P}^{\mathsf{X}} : \mathcal{X} \to [0, 1]$ of X as the *pushforward* of $\mathbb{P}^{\Omega}$ by X. Specifically, for any $x \in X$, $\mathbb{P}^{\mathsf{X}}(x) := \mathbb{P}^{\Omega}(\mathsf{X} \bowtie x)$.

We can alternatively define the pushforward as a Markov kernel product. Let $\mathbb{P}_{\mathsf{X}|\Omega} : \Omega \nrightarrow X$ be the Markov kernel $\mathbb{P}_{\mathsf{X}|\Omega}(x|\omega) \mapsto [\![x = \mathsf{X}(\omega)]\!]$. Then

$$\mathbb{P}^{\mathsf{X}} = \mathbb{P}^{\Omega}\mathbb{P}^{\mathsf{X}|\Omega} \tag{43}$$

Instead of starting with a base measure and pushing it forward to find a probability distribution, we can start with a probability distribution and find base measures that pushforward to that distribution. We call such base measures *completions* of our original distribution.

Definition 2.5 (Completion)**.** Given $\mathsf{X} : \Omega \to X$, an arbitrary probability distribution $\mathbb{P}^{\mathsf{X}} : \mathcal{X} \to [0, 1]$ and a sample space Ω, there are typically a number of base measures $\mathbb{P}^{\Omega}_{\alpha}$ such that $\mathbb{P}^{\mathsf{X}} = \mathbb{P}^{\Omega}_{\alpha}\mathbb{P}^{\mathsf{X}|\Omega}$. We call any $\mathbb{P}^{\Omega}_{\alpha}$ with this property a *completion* of $\mathbb{P}^{\mathsf{X}}$.

Definition 2.6 (Validity)**.** Given $\mathsf{X} : \Omega \to X$, an arbitrary probability distribution $\mathbb{P}^{\mathsf{X}} : \mathcal{X} \to [0, 1]$ and a sample space Ω, $\mathbb{P}^{\mathsf{X}}$ is valid if it has at least one completion.

Probability gap models have types, which indicate which kinds of inserts they accept and what kind of model you will get as a result. Types correspond to random variables; for the following suppose any uppercase sans serif letter X is a random variable $\Omega \to X$:

- An order 0 type X probability gap model is a probability distribution $\mathbb{P}^{\mathsf{X}}$ that has at least one completion to a base measure on Ω

- An order 1 type $\mathsf{X}; \mathsf{Y}$ probability gap model takes a valid order 0 X gap model and returns an order 0 (X, Y) gap model

- An order 2 type $\mathsf{W}; \mathsf{X}; \mathsf{Y}; \mathsf{Z}$ probability gap model takes a valid order 1 $\mathsf{X}; \mathsf{Y}$ gap model and returns an order 1 $\mathsf{W}; (\mathsf{X}, \mathsf{Y}, \mathsf{Z})$ gap model

We will introduce order 1 gap models as *conditional probabilities* and order 2 gap models as *probability 2-combs*.

2.6 Order 1 gaps: conditional probabilities

We use the term *conditional probability* to describe order 1 gap models. A conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ is a valid Markov kernel of the appropriate type.

Definition 2.7 (Valid conditional probability)**.** Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ a *valid conditional probability* $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ is a Markov kernel $X \rightharpoonup Y$ such that it assigns probability 0 to contradictions:

$$\forall y \in Y, x \in X : (\mathsf{X}, \mathsf{Y}) \bowtie (x, y) = \emptyset \implies \left(\mathbb{P}^{\mathsf{Y}|\mathsf{X}}(y|x) = 0 \right) \vee (\mathsf{X} \bowtie (x) = \emptyset) \tag{44}$$

We combine order 0 gap models (probability distributions) with order 1 gap models (conditional probabilities) using the "copy-product" $\odot$.

Definition 2.8 (Copy-product)**.** Given two Markov kernels $\mathbb{K} : X \rightharpoonup Y$ and $\mathbb{L} : Y \times X \rightharpoonup Z$, define the copy-product $\mathbb{K} \odot \mathbb{L} : X \to Y \times Z$ as

$$\mathbb{K} \odot \mathbb{L} := \mathrm{copy}_X (\mathbb{K} \otimes \mathrm{id}_X)(\mathrm{copy}_Y \otimes \mathrm{id}_X)(\mathrm{id}_Y \otimes \mathbb{L}) \tag{45}$$

$$\tag{46}$$

$$\Longleftrightarrow \tag{47}$$

$$(\mathbb{K} \odot \mathbb{L})(y, z|x) = \mathbb{L}(z|y, x)\mathbb{K}(y|x) \tag{48}$$

We say a conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ extends to a probability distribution $\mathbb{P}^{\mathsf{XY}}_\alpha$ via $\mathbb{P}^{\mathsf{X}}_\alpha$ if $\mathbb{P}^{\mathsf{X}|*}_\alpha \odot \mathbb{P}^{\mathsf{Y}|\mathsf{X}*} = \mathbb{P}^{\mathsf{XY}|*}_\alpha$.

If $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ extends to a distribution $\mathbb{Q}^{\mathsf{XY}}$ then we say $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ is a *disintegration* of $\mathbb{Q}^{\mathsf{XY}}$. If the choice of disintegration does not matter, we may use the notation $\overline{\mathbb{Q}}^{\mathsf{Y}|\mathsf{X}}$ to refer to a representative element of this set. The overline is to address an undesirable ambiguity: if we had no overline, we could easily write $\mathbb{Q}^{\mathsf{Y}|\mathsf{X}}$ for any $\mathbb{Q}^{\mathsf{YX}}$ without first checking if it matters which choice we make. Refer to the discussion above about the non-uniqueness of the left hand side of Equation 42 for an example of the problems this can introduce.

Note that our terminology matches a common informal definition of conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ – "a function that give the probability of X given a value of Y". However, the usual formal definition of "the" conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ defines what we would write $\overline{\mathbb{P}}^{\mathsf{Y}|\mathsf{X}}$ given some $\mathbb{P}^{\mathsf{XY}}$; that is, a representative member of the set of disintegrations of $\mathbb{P}^{\mathsf{XY}}$. We choose to call things like $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ "conditional probabilities" because we think it would be confusing to call them anything else. We refer to Hájek (2003) for arguments that our definition is a better match informal usage of this term than the standard definition.

Arbitrary conditional probabilities do not necessarily extend to valid probability distributions. For example, suppose we $X = Y$ and $\mathsf{X} = \mathsf{Y}$ and some $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$

such that, given $x \neq x' \in X$, $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}(x|x') = a > 0$. Take $\mathbb{P}^{\mathsf{X}}_\alpha(x') = 1$ and then $\mathbb{P}^{\mathsf{XY}}_\alpha((x',x)) = a > 0$ also. For every probability model $\mathbb{Q}^{\mathsf{I}} \in \Delta(\Omega)$,

$$\mathbb{Q}^{\mathsf{XY}}(x,x') = \mathbb{Q}^{\mathsf{I}}(\mathsf{X} \bowtie x \cap \mathsf{X} \bowtie x') \tag{49}$$

$$= [\![x = x']\!] \tag{50}$$

Thus $\mathbb{P}^{\mathsf{XY}}_\alpha$ is not valid.

The key justification for our definition of validity is that valid conditional probabilities guarantee that we can extend the conditional probability to a base measure. In fact, Definition 2.7 is equivalent to defining valid conditional probabilities as those that extend to valid probability distributions, given any valid probability distribution (Theorem 2.12). In addition, the defintion of validity itself is equivalent to Definition 2.7 under the identification of probability distributions with trivial conditional probabilities (Theorem 2.9).

Theorem 2.9 (Agreement of validity criteria for conditional probabilities)**.** *Given* $\mathsf{X} : \Omega \to X$*, with* Ω *and* X *discrete, a probability* $\mathbb{P}^{\mathsf{X}}$ *is valid if and only if the conditional probability* $\mathbb{P}^{\mathsf{X}|*} := * \mapsto \mathbb{P}^{\mathsf{X}}$ *is valid.*

Proof. $* \bowtie * = \Omega$ necessarily. Thus validity of $\mathbb{P}^{\mathsf{X}|*}$ means

$$\forall x \in X : \mathsf{X} \bowtie (x) = \emptyset \implies \mathbb{P}^{\mathsf{X}|*}(x|*) = 0 \tag{51}$$

$$= \mathbb{P}^{\mathsf{X}}(x) \tag{52}$$

If: We refer to Ershov (1975) Theorem 2.5 for the proof that Equation 52 is necessary and sufficient for the existence of $\mathbb{P}^{\mathsf{I}}$ such that $\mathbb{P}^{\mathsf{I}}(\mathsf{X}^{-1}(A)) = \mathbb{P}^{\mathsf{X}}(A)$ for all $A \in \mathcal{X}$ when $(\Omega, \mathcal{F})$ and $(X, \mathcal{X})$ are standard measurable. If Ω and X are discrete, then they are standard measurable.

Only if: If $\mathsf{X} \bowtie x = \emptyset$ then $\mathbb{P}^{\mathsf{X}}(x) = \mathbb{P}^{\mathsf{I}}(\emptyset) = 0$. $\qquad \square$

Lemma 2.10 (Valid conditional probabilities are validly extendable)**.** *Given* $(\Omega, \mathcal{F})$*,* $\mathsf{X} : \Omega \to X$*,* $\mathsf{Y} : \Omega \to Y$*,* $\mathsf{Z} : \Omega \to Z$ *and any valid conditional probabilities* $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ *and* $\mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}$*,* $\mathbb{P}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}$ *is also a valid conditional probability.*

Proof. Let $\mathbb{R}^{\mathsf{YZ}|\mathsf{X}} := \mathbb{P}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}$.

We only need to check validity in $x \in \mathsf{X}(\Omega)$, as it is automatically satisfied for other values of X.

For all $x \in \mathsf{X}(\Omega)$, $\mathsf{X} \bowtie x \cap \mathsf{Y} \bowtie y = \emptyset$, $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}(y|x) = 0$ by validity. Thus

$$\mathbb{R}^{\mathsf{YZ}|\mathsf{X}}(y,z|x) = \mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}(z|y,x)\mathbb{P}^{\mathsf{Y}|\mathsf{X}}(y|x) \tag{53}$$

$$\leq \mathbb{P}^{\mathsf{Y}|\mathsf{X}}(y|x) \tag{54}$$

$$= 0 \tag{55}$$

For all $(x,y) \in (\mathsf{X},\mathsf{Y})(\Omega)$, $z \in Z$ such that $(\mathsf{X},\mathsf{Y},\mathsf{Z}) \bowtie (x,y,z) = \emptyset$, $\mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}(z|y,x) = 0$ by validity. Thus for any such (x,y,z):

$$\mathbb{R}^{\mathsf{YZ}|\mathsf{X}}(y,z|x) = \mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}(z|y,x)\mathbb{P}^{\mathsf{Y}|\mathsf{X}}(y|x) \tag{56}$$

$$= 0 \tag{57}$$

$$\square$$

Corollary 2.11 (Valid conditional probability is validly extendable to a probability distribution)**.** *Given Ω, $\mathsf{U} : \Omega \to U$, $\mathsf{W} : \Omega \to W$ and a valid conditional probability $\mathbb{T}^{\mathsf{W}|\mathsf{U}}$, then for any valid conditional probability $\mathbb{V}^{\mathsf{U}}$, $\mathbb{V}^{\mathsf{U}} \odot \mathbb{T}^{\mathsf{W}|\mathsf{U}}$ is a valid probability distribution.*

Proof. Applying Lemma 2.10 choosing $\mathsf{X} = *$, $\mathsf{Y} = \mathsf{U}$, $\mathsf{Z} = \mathsf{W}$ and $\mathbb{P}^{\mathsf{Y}|\mathsf{X}} = \mathbb{V}^{\mathsf{U}|*}$ and $\mathbb{Q}^{\mathsf{Z}|\mathsf{YX}} = \mathbb{T}^{\mathsf{W}|\mathsf{U}*}$ we have $\mathbb{R}^{WU|*} := \mathbb{V}^{\mathsf{U}|*} \odot \mathbb{T}^{\mathsf{W}|\mathsf{U}*}$ is a valid conditional probability. Then $\mathbb{R}^{\mathsf{WU}} \cong \mathbb{R}^{\mathsf{WU}|*}$ is valid by Theorem 2.9. $\qquad\square$

Theorem 2.12 (Validity of conditional probabilities)**.** *Suppose we have Ω, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, with Ω, X, Y discrete. A conditional probability $\mathbb{T}^{\mathsf{Y}|\mathsf{X}}$ is valid if and only if for all valid probability distributions $\mathbb{V}^{\mathsf{X}}$, $\mathbb{V}^{\mathsf{X}} \odot \mathbb{T}^{\mathsf{Y}|\mathsf{X}}$ is a valid probability distribution.*

Proof. If: this follows directly from Lemma 2.10.

Only if: suppose $\mathbb{T}^{\mathsf{Y}|\mathsf{X}}$ is invalid. Then there is some $x \in X$, $y \in Y$ such that $\mathsf{X} \bowtie (x) \neq \emptyset$, $(\mathsf{X}, \mathsf{Y}) \bowtie (x, y) = \emptyset$ and $\mathbb{T}^{\mathsf{Y}|\mathsf{X}}(y|x) > 0$. Choose $\mathbb{V}^{\mathsf{X}}$ such that $\mathbb{V}^{\mathsf{X}}(\{x\}) = 1$; this is possible due to standard measurability and valid due to $\mathsf{X}^{-1}(x) \neq \emptyset$. Then

$$(\mathbb{V}^{\mathsf{X}} \odot \mathbb{T}^{\mathsf{Y}|\mathsf{X}})(x, y) = \mathbb{T}^{\mathsf{Y}|\mathsf{X}}(y|x)\mathbb{V}^{\mathsf{X}}(x) \tag{58}$$

$$= \mathbb{T}^{\mathsf{Y}|\mathsf{X}}(y|x) \tag{59}$$

$$> 0 \tag{60}$$

Hence $\mathbb{V}^{\mathsf{X}} \odot \mathbb{T}^{\mathsf{Y}|\mathsf{X}}$ is invalid. $\qquad\square$

Given any two conditional probabilities $\mathbb{T}^{\mathsf{Y}|\mathsf{X}}$, $\mathbb{U}^{\mathsf{Y}|\mathsf{X}}$ such that there is some $x \in \mathsf{X}(\Omega)$, $A \in \mathcal{Y}$ for which $\mathbb{T}^{\mathsf{Y}|\mathsf{X}}(A|x) \neq \mathbb{U}^{\mathsf{Y}|\mathsf{X}}(A|x)$, there exists some $\mathbb{Q}^{\mathsf{X}}$ such that $\mathbb{Q}^{\mathsf{X}} \odot \mathbb{T}^{\mathsf{Y}|\mathsf{X}} \neq \mathbb{Q}^{\mathsf{X}} \odot \mathbb{U}^{\mathsf{Y}|\mathsf{X}}$. We can see this by letting $\mathbb{Q}^{\mathsf{X}}$ assign probability 1 to some value of x on which $\mathbb{T}^{\mathsf{Y}|\mathsf{X}}$ and $\mathbb{U}^{\mathsf{Y}|\mathsf{X}}$ differ. Thus if we are using conditional probability as a probability model with a gap, uniqueness requires that the conditional probability be represented by a Markov kernel that is unique up to the set of impossible values of the conditioning variable. In contrast, conditional probability derived from a standard probability model may be represented by a Markov kernel unique up to a measure zero set, which is a strictly weaker condition because the set of impossible values must be given measure 0 by any valid probability distribution.

We finally note that a conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ along with the copy-product $\odot$ corresponds precisely to the set of insert compatible mixture homomorphisms.

Definition 2.13 (Marginalisation)**.**

Lemma 2.14 (Marginal probabilities are unique compatible submodels)**.**

Definition 2.15 (Delta notation)**.**

Definition 2.16 (Order 1 insert compatibility)**.** Given a function $f : \Delta(X) \to \Delta(X \times Y)$, f is *insert-compatible* if

$$\tag{61}$$

Definition 2.17 (Mixture homomorphic)**.** Given a function $f : \Delta(X) \to \Delta(X \times Y)$, f is *mixture homomorphic* if

$$f(a\mathbb{P}_1^{\mathsf{X}} + (1-a)\mathbb{P}_2^{\mathsf{X}}) = af(\mathbb{P}_1^{\mathsf{X}}) + (1-a)f(\mathbb{P}_2^{\mathsf{X}}) \tag{62}$$

Proposition: Every insert-compatible, mixture homomorphic function from $\Delta(X) \to \Delta(X \times Y)$ can be identified with a Markov kernel $X \twoheadrightarrow X \times Y$.

Proof sketch: We can write any probability measure over a discrete space as a mixture of delta-measures. Define $\mathbb{K}(x, y | x') := f(\delta_{x'})(x, y)$. Then

$$f(\mu)(x, y) = f(\sum_{i \in X} \mu(i)\delta_i)(x, y) \tag{63}$$

$$= \sum_{i \in X} \mu(i)f(\delta_i)(x, y) \tag{64}$$

$$= \sum_{i \in X} \mu(i)\mathbb{K}(x, y | i) \tag{65}$$

$$= (\mu\mathbb{K})(x, y) \tag{66}$$

Due to insert compatibility, we also require for some $\mathbb{L} \in \Delta(Y)$

$$\sum_y f(\delta_{x'})(x, y) = [\![x = x']\!] \tag{67}$$

$$\implies f(\delta_{x'})(x, y) = [\![x = x']\!]\mathbb{L}(y|x) \tag{68}$$

Thus

$$f(\mu) = \mu \odot \mathbb{L} \tag{69}$$

2.7 Order 2 gaps: probability combs

An order-2 gap model can be extended with an order-1 gap model to yield an order-1 gap model. Order-2 gap models are *probability 2-combs* (Chiribella et al., 2008; Jacobs et al., 2019).

Definition 2.18 (Probability 2-comb)**.** Given $\mathsf{W} : \Omega \to W$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$, a probability 2-comb $\mathbb{P}^{\mathsf{X}|\mathsf{W}\square\mathsf{Z}|\mathsf{Y}} : W \times Y \twoheadrightarrow X \times Z$ is a Markov kernel such that for some $\mathbb{P}^{\mathsf{X}|\mathsf{W}} : W \twoheadrightarrow X$

$$\tag{70}$$

Coupled with the validity conditions

17

1. $\mathbb{P}^{X|W}$ is valid as a conditional probability

2. $(W, X, Y, Z) \bowtie (w, x, y, z) = \emptyset$ implies $\mathbb{P}^{X|W\square Z|Y}(x, z|w, y) = 0$ or $(W, X, Y) \bowtie (w, x, y) = \emptyset$

With discrete sets, and in general wherever we have kernel disintegrations, there exists some $\overline{\mathbb{P}}^{Z|WXY} : W \times X \times Y \rightarrow Z$ (Lemma 2.24) such that

$$\mathbb{P}^{X|W\square Z|Y} = \qquad \qquad \qquad \qquad \qquad \qquad \qquad (71)$$

We define the operation insert that takes a 2-comb and a conditional probability and returns a conditional probability.

$$\text{insert}(\mathbb{P}_\alpha^{Y|XW}, \mathbb{P}^{X|W\square Z|Y}) = \mathbb{P}^{X|W} \odot \mathbb{P}_\alpha^{Y|XW} \odot \overline{\mathbb{P}}^{Z|WXY} \qquad (72)$$

We can depict this operation graphically in a somewhat informal way as "inserting" $\mathbb{P}_\alpha^{Y|XW}$ into $\mathbb{P}^{X|W\square Z|Y}$:

$$\text{Insert}\left(\qquad \qquad , \qquad \qquad \right) \qquad (73)$$

$$= \qquad \qquad \qquad \qquad \qquad \qquad \qquad (74)$$

Just as in Section 2.6, a key concern for probability 2-combs is *valid extendability*. The insert operation we define for 2-combs is uniquely defined by Equation 72 (Theorem 2.19, and yields a valid conditional probability given any valid inputs (Theorem 2.20).

Theorem 2.19 (Equivalence of 2-comb disintegrations). *Given a 2-comb $\mathbb{P}^{X|W\square Z|Y}$ and any two disintegrations $\overline{\mathbb{P}}_1^{Z|WXY}$, $\overline{\mathbb{P}}_2^{Z|WXY}$, for all valid extensions $\mathbb{P}_\alpha^{Y|XW}$*

$$\mathbb{P}^{X|W} \odot \mathbb{P}_\alpha^{Y|XW} \odot \overline{\mathbb{P}}_1^{Z|WXY} = \mathbb{P}^{X|W} \odot \mathbb{P}_\alpha^{Y|XW} \odot \overline{\mathbb{P}}_2^{Z|WXY} \qquad (75)$$

Proof. For any w, x, y, z

$$(\mathbb{P}^{\mathsf{X}|\mathsf{W}} \odot \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{XW}} \odot \overline{\mathbb{P}}_1^{\mathsf{Z}|\mathsf{WXY}})(x,y,z|w) = \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}}(y|w,x)\mathbb{P}^{\mathsf{X}|\mathsf{W}}(x|w)\overline{\mathbb{P}}_1^{\mathsf{Z}|\mathsf{XYW}}(z|x,y,w) \tag{76}$$

$$= \mathbb{P}^{\mathsf{X}|\mathsf{W}\square\mathsf{Z}|\mathsf{Y}}(x,z|w,y) \tag{77}$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}}(y|w,x)\mathbb{P}^{\mathsf{X}|\mathsf{W}}(x|w)\overline{\mathbb{P}}_2^{\mathsf{Z}|\mathsf{XYW}}(z|x,y,w) \tag{78}$$

$\square$

Theorem 2.20 (Extension of valid probability 2-combs). *Given Ω, $\mathsf{W} : \Omega \to W$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and $\mathsf{Z} : \Omega \to Z$, a probability 2-comb $\mathbb{P}^{\mathsf{X}|\mathsf{W}\square\mathsf{Z}|\mathsf{Y}}$ is valid if and only if $\mathrm{insert}(\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}}, \mathbb{P}^{\mathsf{X}|\mathsf{W}\square\mathsf{Z}|\mathsf{Y}})$ is valid for all valid $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}}$.*

Proof. Only if:
  Note that

$$\mathbb{P}_\alpha^{\mathsf{XYZ}|\mathsf{W}} := \mathrm{insert}(\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}}, \mathbb{P}^{\mathsf{X}|\mathsf{W}\square\mathsf{Z}|\mathsf{Y}}) \tag{79}$$

$$\mathbb{P}_\alpha^{\mathsf{XYZ}|\mathsf{W}}(xyz|w) = \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}}(y|w,x)\mathbb{P}^{\mathsf{X}|\mathsf{W}}(x|w)\overline{\mathbb{P}}^{\mathsf{Z}|\mathsf{XYW}}(z|x,y,w) \tag{80}$$

$$= \mathbb{P}^{\mathsf{X}|\mathsf{W}\square\mathsf{Z}|\mathsf{Y}}(x,z|w,y)\mathbb{P}_\alpha(y|x) \tag{81}$$

Suppose $\mathbb{P}^{\mathsf{X}|\mathsf{W}\square\mathsf{Z}|\mathsf{Y}}$ is valid. If $(\mathsf{W}, \mathsf{X}, \mathsf{Y}, \mathsf{Z}) \bowtie (w,x,y,z) = \emptyset$ then either $\mathbb{P}^{\mathsf{X}|\mathsf{W}\square\mathsf{Z}|\mathsf{Y}}(x,z|w,y) = 0$ and hence $\mathbb{P}_\alpha^{\mathsf{XYZ}|\mathsf{W}}(xyz|w) = 0$ or $(\mathsf{W}, \mathsf{X}, \mathsf{Y}) \bowtie (w,x,y) = \emptyset$.

If $(\mathsf{W}, \mathsf{X}, \mathsf{Y}) \bowtie (w,x,y) = \emptyset$ then either $(\mathsf{W}, \mathsf{X}) \bowtie (w,x) \neq \emptyset$ and by validity $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}}(y|w,x) = 0$ and so $\mathbb{P}_\alpha^{\mathsf{XYZ}|\mathsf{W}}(xyz|w) = 0$ or $(\mathsf{W}, \mathsf{X}) \bowtie (w,x) = \emptyset$.

If $(\mathsf{W}, \mathsf{X}) \bowtie (w,x) = \emptyset$ then either $\mathsf{W} \bowtie w \neq \emptyset$ and by validity $\mathbb{P}^{\mathsf{X}|\mathsf{W}}(x|w) = 0$ and so $\mathbb{P}_\alpha^{\mathsf{XYZ}|\mathsf{W}}(xyz|w) = 0$ or $\mathsf{W} \bowtie w = \emptyset$, in which case $\mathbb{P}_\alpha^{\mathsf{XYZ}|\mathsf{W}}(xyz|w)$ may take any value.

If: Suppose $\mathbb{P}^{\mathsf{X}|\mathsf{W}\square\mathsf{Z}|\mathsf{Y}}$ is invalid. Then either $\mathbb{P}^{\mathsf{X}|\mathsf{W}}$ is invalid or $\mathbb{P}^{\mathsf{X}|\mathsf{W}\square\mathsf{Z}|\mathsf{Y}}(x,z|w,y) > 0$ on some (w,x,y,z) such that $(\mathsf{W}, \mathsf{X}, \mathsf{Y}, \mathsf{Z}) \bowtie (w,x,y,z) = \emptyset$ and $(\mathsf{W}, \mathsf{X}, \mathsf{Y}) \bowtie (w,x,y) \neq \emptyset$.

Suppose $\mathbb{P}^{\mathsf{X}|\mathsf{W}}$ is invalid. Then

$$\mathbb{P}_\alpha^{\mathsf{X}|\mathsf{W}}(x|w) = \sum_{y \in Y, z \in Z} \mathbb{P}_\alpha^{\mathsf{XYZ}|\mathsf{W}}(xyz|w) \tag{82}$$

$$= \mathbb{P}^{\mathsf{X}|\mathsf{W}}(x|w) \tag{83}$$

Thus $\mathbb{P}_\alpha^{\mathsf{X}|\mathsf{W}}(x|w)$ is invalid and therefore so too is $\mathbb{P}_\alpha^{\mathsf{XYZ}|\mathsf{W}}$.

Suppose we have some (w,x,y,z) such that $(\mathsf{W}, \mathsf{X}, \mathsf{Y}, \mathsf{Z}) \bowtie (w,x,y,z) = \emptyset$, $(\mathsf{W}, \mathsf{X}, \mathsf{Y}) \bowtie (w,x,y) \neq \emptyset$ and $\mathbb{P}^{\mathsf{X}|\mathsf{W}\square\mathsf{Z}|\mathsf{Y}}(x,z|w,y) > 0$.

By supposition, there is a valid $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}}$ such that $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}}(y|w,x) = 1$. Then

$$\mathbb{P}_\alpha^{\mathsf{XYZ|W}}(xyz|w) = \mathbb{P}^{\mathsf{X|W\square Z|Y}}(x,z|w,y)\mathbb{P}_\alpha(y|x) \tag{84}$$
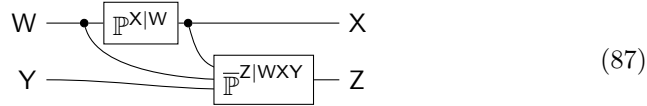
$$= \mathbb{P}^{\mathsf{X|W\square Z|Y}}(x,z|w,y) \tag{85}$$

$$> 0 \tag{86}$$

So $\mathbb{P}_\alpha^{\mathsf{XYZ|W}}$ is invalid. $\qquad\square$

It is also the case that if we combine two valid conditional probabilities to form a 2-comb, the result is a valid 2-comb.

Theorem 2.21 (Valid conditional probabilities combine to a valid 2-comb)**.** *Given valid conditional probabilities* $\mathbb{P}^{\mathsf{X|W}}$ *and* $\mathbb{P}^{\mathsf{Z|WXY}}$

$$(87)$$

Is a valid 2-comb.

Proof. Validity of $\mathbb{P}^{\mathsf{X|W}}$ is by assumption, and validity of $\mathbb{P}^{\mathsf{Z|WXY}}$ means $(\mathsf{W,X,Y,Z}) \bowtie (w,x,y,z) = \emptyset$ implies $\mathbb{P}^{\mathsf{Z|WXY}}(z|w,x,y) = 0$ or $(\mathsf{W,X,Y}) \bowtie (w,x,y) = \emptyset$. This in turn implies $\mathbb{P}^{\mathsf{Z|WXY}}(z|w,x,y) = 0$ or $(\mathsf{W,X,Y}) \bowtie (w,x,y) = \emptyset$. $\qquad\square$

2.8 Revisiting truncated factorisation

We have established that, at least for discrete sets, probability 2-combs are a general kind of object that maps a conditional probability to a conditional probability. Thus they model choices that correspond to conditional probabilities, rather than probability distributions. We initially considered that a truncated factorisation might be a conditional probability – that is, it could be used to model choices that correspond to probability distributions of the choice variable U that we introduced. If we instead consider the possibility that the model of "interventions" that Equation 42 is suggesting is actually a probability 2-comb, we find that both of the problems we identified are automatically solved.

Suppose we have a 2-comb $\mathbb{P}^{\mathsf{Z\square Y|X}}$. Then we have some $\mathbb{P}^{\mathsf{Z}}$, $\overline{\mathbb{P}}^{\mathsf{Y|XZ}}$ unique up to equivalence such that

$$\mathbb{P}^{\mathsf{Z\square Y|X}}(z,y|x) = \overline{\mathbb{P}}^{\mathsf{Y|XZ}}(y|x,z)\mathbb{P}^{\mathsf{Z}}(z) \tag{88}$$

By Lemma 2.23, we also have

$$\mathbb{P}^{\mathsf{Z\square XY|X}}(z,x',y|x) = \overline{\mathbb{P}}^{\mathsf{Y|XZ}}(y|x,z)\mathbb{P}^{\mathsf{Z}}(z)[\![x=x']\!] \tag{89}$$

Compare this to Equation 42, which we reproduce here for convenience:

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

These equations look almost identical, and suggest an alternative approach to modelling intervention. Instead of *defining* an interventional conditional probability via truncated factorisation, suppose we have an interventional model that is a probability 2-comb, and that satisfies

$$\mathbb{P}^{\mathsf{XYZ}}_{\mathrm{obs}} = \mathrm{insert}(\mathbb{P}^{\mathsf{X}|\mathsf{Z}}_{\mathrm{obs}}, \mathbb{P}^{\mathsf{Z}\square\mathsf{XY}|\mathsf{X}}) \tag{90}$$

For some observational insert $\mathbb{P}^{\mathsf{X}|\mathsf{Z}}_{\mathrm{obs}}$.

Under this view, we cannot necessarily derive $\mathbb{P}^{\mathsf{Z}\square\mathsf{XY}|\mathsf{X}}$ from the observational probability distribution $\mathbb{P}^{\mathsf{XYZ}}_{\mathrm{obs}}$, as we cannot guarantee that any particular $\overline{\mathbb{P}}^{\mathsf{Y}|\mathsf{XZ}}_{\mathrm{obs}}$ is a representative of $\overline{\mathbb{P}}^{\mathsf{Y}|\mathsf{XZ}}$. This is as it should be – if we have meaningfully different choices of $\overline{\mathbb{P}}^{\mathsf{Y}|\mathsf{XZ}}_{\mathrm{obs}}$ arising from measure zero sets, we cannot decide which one will model interventions without more information. The 2-comb $\mathbb{P}^{\mathsf{Z}\square\mathsf{XY}|\mathsf{X}}$ itself stands for a unique map from interventions to probability distributions, rather than the potentially large set of different maps of this type.

The second issue we raised was the nonexistence of $\mathbb{P}^{\mathsf{XYZ}}_x$ if, for example, $\mathsf{X} = f\circ\mathsf{Z}$. Because $\mathbb{P}^{\mathsf{Z}}_{\mathrm{obs}}$ and $\overline{\mathbb{P}}^{\mathsf{Y}|\mathsf{XZ}}_{\mathrm{obs}}$ are both valid, Lemma 2.21 ensures that any 2-comb defined by any choice of these conditional probabilities is also valid, and so Equation 89 defines a valid object. Because we have a 2-comb which is extended by *conditional probabilities* rather than by probability distributions, and these conditional probabilities *must be valid*, we can validly extend $\mathbb{P}^{\mathsf{Z}\square\mathsf{XY}|\mathsf{X}}$ even when $\mathsf{X} = f \circ \mathsf{Z}$ – it's just that validity requires that any extending conditional must be such that $\mathbb{P}^{\mathsf{X}|\mathsf{Z}}_\alpha(x|z) = [\![x = f(z)]\!]$.

> So: 2-combs do what we want truncated factorisation to do without breaking down in edge cases, except maybe when we go to uncountably infinite sets.

2.9 Useful results

2.9.1 Repeated variables

Lemmas 2.22 and 2.23 establish that models of repeated variables must connect the repetitions with a copy map.

Lemma 2.22 (Output copies of the same variable are identical)**.** *For any Ω, $\mathsf{X}, \mathsf{Y}, \mathsf{Z}$ random variables on Ω and conditional probability $\mathbb{K}^{\mathsf{YZ}|\mathsf{X}}$, there is a conditional probability $\mathbb{K}^{\mathsf{YYZ}|\mathsf{X}}$ unique up to impossible values of X such that*

$$\mathsf{X} \!-\!\!\boxed{\mathbb{K}^{\mathsf{YYZ}|\mathsf{X}}} \genfrac{}{}{0pt}{}{\overset{*}{=\!=}}{\substack{\mathsf{Y}\\ \mathsf{Z}}} = \mathbb{K}^{\mathsf{YZ}|\mathsf{X}} \tag{91}$$

and it is given by

$$\mathbb{K}^{\mathsf{YYZ|X}} = \quad \boxed{\text{diagram}} \tag{92}$$

$$\Longleftrightarrow \tag{93}$$

$$\mathbb{K}^{\mathsf{YYZ|X}}(y, y', z|x) = [\![y = y']\!]\mathbb{K}^{\mathsf{YZ|X}}(y, z|x) \tag{94}$$

$$\tag{95}$$

Proof. If we have a valid $\mathbb{K}^{\mathsf{YYZ|X}}$, it must be the pushforward of $(\mathsf{Y}, \mathsf{Y}, \mathsf{Z})$ under some $\mathbb{K}^{\mathsf{I|X}}$. Furthermore, $\mathbb{K}^{\mathsf{YZ|X}}$ must be the pushforward of $(*, \mathsf{Y}, \mathsf{Z}) \cong (\mathsf{Y}, \mathsf{Z})$ under the same $\mathbb{K}^{\mathsf{I|X}}$.

For any $x \in \mathsf{X}(\Omega)$, validity requires $(\mathsf{X}, \mathsf{Y}, \mathsf{Y}, \mathsf{Z}) \bowtie (x, y, y', z) = \emptyset \implies \mathbb{K}^{\mathsf{YYZ|X}}(y, y', z|x) = 0$. Clearly, whenever $y \neq y'$, $\mathbb{K}^{\mathsf{YYZ|X}}(y, y', z|x) = 0$. Because $\mathbb{K}^{\mathsf{YYZ|X}}$ is a Markov kernel, there is some $\mathbb{L} : X \rightharpoonup X \times Z$ such that

$$\mathbb{K}^{\mathsf{YYZ|X}}(y, y', z|x) = [\![y = y']\!]\mathbb{L}(y, z|x) \tag{96}$$

$$\tag{97}$$

But then

$$\mathbb{K}^{\mathsf{YZ|X}}(y, z|x) = \sum_{y' \in Y} \mathbb{K}^{\mathsf{YYZ|X}}(y, y', z|x) \tag{98}$$

$$= \mathbb{L}(y, z|x) \tag{99}$$

$$\tag{100}$$

$$\square$$

Lemma 2.23 (Copies shared between input and output are identical)**.** *For any* $\mathbb{K} : (\mathsf{X}, \mathsf{Y}) \rightharpoonup (\mathsf{X}, \mathsf{Z})$, $\mathbb{K}$ *is a model iff there exists some* $\mathbb{L} : (\mathsf{X}, \mathsf{Y}) \rightharpoonup \mathsf{Z}$ *such that*

$$\mathbb{K} = \quad \boxed{\text{diagram}} \tag{101}$$

$$\Longleftrightarrow \tag{102}$$

$$\mathbb{K}'^{x',z}_{x,y} = [\![x = x']\!]\mathbb{L}^{z}_{/x,y} \tag{103}$$

For any $\Omega, \mathsf{X}, \mathsf{Y}, \mathsf{Z}$ *random variables on* Ω *and conditional probability* $\mathbb{K}^{\mathsf{Z|XY}}$, *there is a conditional probability* $\mathbb{K}^{\mathsf{XZ|XY}}$ *unique up to impossible values of* (X, Y) *such that*

$$\boxed{\text{diagram}} = \mathbb{K}^{\mathsf{XZ|XY}} \tag{104}$$

22

and it is given by

$$\mathbb{K}^{\mathsf{XZ|XY}} = \mathsf{X} \longrightarrow \boxed{\mathbb{K}^{\mathsf{YZ|X}}} \begin{array}{l} \mathsf{Y} \\ \mathsf{Y} \\ \mathsf{Z} \end{array} \tag{105}$$

$$\Longleftrightarrow \tag{106}$$

$$\mathbb{K}^{\mathsf{XZ|XY}}(x, z|x', y) = [\![x = x']\!] \mathbb{K}^{\mathsf{Z|XY}}(z|x', y) \tag{107}$$

$$\tag{108}$$

Proof. If we have a valid $\mathbb{K}^{\mathsf{XZ|XY}}$, it must be the pushforward of (X, Z) under some $\mathbb{K}^{\mathsf{I|XY}}$. Furthermore, $\mathbb{K}^{\mathsf{Z|XY}}$ must be the pushforward of $(*, \mathsf{Z}) \cong (\mathsf{Z})$ under the same $\mathbb{K}^{\mathsf{I|X}}$.

For any $(x, y) \in (\mathsf{X}, \mathsf{Y})(\Omega)$, validity requires $(\mathsf{X}, \mathsf{Y}, \mathsf{X}, \mathsf{Z}) \bowtie (x, y, x', z) = \emptyset \implies \mathbb{K}^{\mathsf{XZ|XY}}(x', z|x, y) = 0$. Clearly, whenever $x \neq x'$, $\mathbb{K}^{\mathsf{XZ|XY}}(x', z|x, y) = 0$. Because $\mathbb{K}^{\mathsf{XZ|XY}}$ is a Markov kernel, there is some $\mathbb{L} : X \times Y \rightarrow Z$ such that

$$\mathbb{K}^{\mathsf{XZ|XY}}(x', z|x, y) = 0 = [\![x = x']\!] \mathbb{L}(z|x, y) \tag{109}$$

$$\tag{110}$$

But then

$$\mathbb{K}^{\mathsf{Z|XY}}(y, z|x) = \sum_{x' \in X} \mathbb{K}^{\mathsf{XZ|XY}}(x', z|x, y) \tag{111}$$

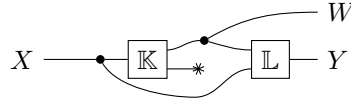$$= \mathbb{L}(z|x, y) \tag{112}$$

$$\tag{113}$$

$$\square$$

2.9.2 Disintegrations

Lemma 2.24 (Disintegration existence)**.** *For any model* $\mathbb{K} : X \rightarrow W \times Y$ *where* X, W, Y *are discrete, there exists* $\mathbb{L} : W \times X \rightarrow Y$ *such that*

$$\mathbb{K} = \tag{114}$$

Proof. Consider any Markov kernel $\mathbb{L} : W \times X \rightarrow Y$ with the property

$$\mathbb{L}(y|w, x) = \frac{\mathbb{K}(w, y|x)}{\sum_{y \in Y} \mathbb{K}(w, y|x)} \qquad \forall x, w : \text{ the denominator is positive} \tag{115}$$

Then

$$\sum_{y\in Y}\mathbb{K}(w,y|x)\mathbb{L}(y|w,x) = \sum_{y\in Y}\mathbb{K}(w,y|x)\frac{\mathbb{K}(w,y|x)}{\sum_{y\in Y}\mathbb{K}(w,y|x)} \quad \text{if } \sum_{y\in Y}\mathbb{K}(w,y|x) > 0 \tag{116}$$

$$= \mathbb{K}(w,y|x) \quad\quad\quad \text{if } \sum_{y\in Y}\mathbb{K}(w,y|x) > 0 \tag{117}$$

$$= 0 \quad\quad\quad\quad\quad \text{otherwise} \tag{118}$$

$$= \mathbb{K}(w,y|x) \quad\quad\quad\quad \text{otherwise} \tag{119}$$

In general there are many indexed Markov kernels that satisfy this. $\qquad\square$

Lemma 2.25 (Valid disintegration choice). *Given valid $\mathbb{K}^{WY|X}$, there exists a valid conditional probability $\overline{\mathbb{L}}^{Y|WX}$ such that*

$$\mathbb{K}^{WY|X} = \mathbb{K}^{W|X} \odot \mathbb{L}^{Y|WX} \tag{120}$$

Proof. From Lemma 2.24, we have the existence of some $\overline{\mathbb{L}}^{Y|WX} : W \times X \to Y$. We need to check that $\overline{\mathbb{L}}^{Y|WX}$ can be chosen so that it is valid. By validity of $\mathbb{K}^{W,Y|X}$, $w \in W(\Omega)$ and $(X,W,Y) \bowtie (x,w,y) = \emptyset \implies \mathbb{K}^{W,Y|X} = 0$, so we only need to check for (w,x,y) such that $\mathbb{K}^{W,Y|X}(w,y|x) = 0$. For all x,y such that $\mathbb{K}^{Y|X}(y|x)$ is positive, we have $\mathbb{K}^{W,Y|X}(w,y|x) = 0 \implies \overline{\mathbb{L}}^{Y|WX}(y|w,x) = 0$. Furthermore, where $\mathbb{K}^{W|X}(w|x) = 0$, we either have $(W,X) \bowtie (w,x) = \emptyset$ or we can choose some $\omega \in (W,X) \bowtie (w,x)$ and let $\overline{\mathbb{L}}^{Y|WX}(Y(\omega)|w,x) = 1$. $\qquad\square$

2.9.3 Conditional independence

Conditional independence has a familiar definition in probability models. For models with probability gaps, we propose the convention that a conditional independence statement for such a model is a claim that conditional independence holds for the whole set of probability models it is associated with. That is, B is conditionally independent of A given C just when that conditional independence holds for all possible extensions to a standard probability model. This convention could also be applied to other kinds of incomplete probability models beyond the hierarchy proposed here.

Definition 2.26 (Order 0 conditional independence). For a *probability model* $\mathbb{P}^I$ and variables A, B, Z, we say B is conditionally independent of A given C,

written $B \perp\!\!\!\perp_\mathbb{P} A|C$, if

$$\mathbb{P}^{\mathsf{ABC}} = \vcenter{\hbox{}} \qquad (121)$$

For any $\overline{\mathbb{P}}^{\mathsf{B|C}}$ and $\overline{\mathbb{P}}^{\mathsf{A|C}}$. Cho and Jacobs (2019) have shown that this definition coincides with the standard notion of conditional independence. In particular, it satisfies the *semi-graphoid axioms*

1. Symmetry: $A \perp\!\!\!\perp_\mathbb{P} B|C$ iff $B \perp\!\!\!\perp_\mathbb{P} A|C$

2. Decomposition: $A \perp\!\!\!\perp_\mathbb{P} (B,C)|W$ implies $A \perp\!\!\!\perp_\mathbb{P} Y|W$ and $A \perp\!\!\!\perp_\mathbb{P} C|W$

3. Weak union: $A \perp\!\!\!\perp_\mathbb{P} (B,C)|W$ implies $A \perp\!\!\!\perp_\mathbb{P} B|(C,W)$

4. Contraction: $A \perp\!\!\!\perp_\mathbb{P} C|W$ and $A \perp\!\!\!\perp_\mathbb{P} B|(C,W)$ implies $A \perp\!\!\!\perp_\mathbb{P} (B,C)|W$

Conditional independence $A \perp\!\!\!\perp_\mathbb{P} B|C$ holds for an arbitrary probability distribution $\mathbb{P}^{\mathsf{W}}$ if it holds for all probability models $\mathbb{P}^{\mathsf{I}}_\alpha$ that pushforward to $\mathbb{P}^{\mathsf{W}}$. This might happen when, for example, $(A,B,C) = W$. Thus we define order 0 conditional independence with respct to an arbitrary probability distribution as well.

Definition 2.27 (Order 1 conditional independence)**.** For a *conditional probability model* $\mathbb{P}^{\mathsf{V|W}}$, define $\mathbb{P}^{\mathsf{VW}}_\beta := \mathbb{P}^{\mathsf{W}}_\beta \odot \mathbb{P}^{\mathsf{V|W}}$. Then say B is order 1 conditionally independent of A given C, written $B \perp\!\!\!\perp^1_\mathbb{P} A|C$ if for all β, $B \perp\!\!\!\perp_{\mathbb{P}_\beta} A|C$.

In the case of a conditional probability model $\mathbb{P}^{\mathsf{V|W}}$, we can view $\mathbb{P}^{\mathsf{V|W}}$ as a collection of probability distributions $\{\mathbb{P}^{\mathsf{V}}_w := A \mapsto \mathbb{P}^{\mathsf{V|W}}(A|w)|w \in W\}$. We can then consider W to be what Constantinou and Dawid (2017) call "the regime indicator" and V what they call a "stochastic variable". That paper introduces a notion of *extended conditional independence*, which can be applied to models containing combination of stochastic variables and regime variables like $\mathbb{P}^{\mathsf{V|W}}$. Theorem 4.4 of that work proves the following claim:

Theorem 2.28. *Let* $A^* = A \circ V$, $B^* = B \circ V$, $C^* = C \circ V$ *((A, B, C) are $\mathcal{V}$-measurable) and* $D^* = D \circ W, E^* = E \circ W$ *where W is discrete and $W = (D^*, E^*)$. In addition, let $\mathbb{P}^{\mathsf{W}}_\alpha$ be some probability distribution on W such that $w \in \mathsf{W}(\Omega) \implies \mathbb{P}^{\mathsf{W}}_\alpha(w) > 0$. Then, denoting extended conditional independence with $\perp\!\!\!\perp_{\mathbb{P},ext}$ and $\mathbb{P}^{\mathsf{VW}}_\alpha := \mathbb{P}^{\mathsf{W}}_\alpha \odot \mathbb{P}^{\mathsf{V|W}}$*

$$A \perp\!\!\!\perp_{\mathbb{P},ext} (B,D)|(C,E) \iff A^* \perp\!\!\!\perp_{\mathbb{P}_\alpha} (B^*,D^*)|(C^*,E^*) \qquad (122)$$

Where $\perp\!\!\!\perp_{\mathbb{P}_\alpha}$ is order 0 conditional independence.

This result implies a close relationship between order 1 condtional independence and extended conditional independence.

Theorem 2.29. *Let* $A^* = A \circ V$, $B^* = B \circ V$, $C^* = C \circ V$ *((A, B, C) are* $\mathcal{V}$*-measurable) and* $D^* = D \circ W$, $E^* = E \circ W$ *where* V, W *are discrete and* $W = (D^*, E^*)$. *Then letting* $\mathbb{P}_\alpha^{VW} := \mathbb{P}_\alpha^W \odot \mathbb{P}^{V|W}$

$$A \perp\!\!\!\perp^1_{\mathbb{P},ext} (B,D)|(C,E) \iff A^* \perp\!\!\!\perp_{\mathbb{P}} (B^*, D^*)|(C^*, E^*) \tag{123}$$

Proof. If:

By assumption, $A^* \perp\!\!\!\perp_{\mathbb{P}_\alpha} (B^*, D^*)|(C^*, E^*)$ for all $\mathbb{P}_\alpha^{D^*E^*}$. In particular, this holds for some $\mathbb{P}_\alpha^{D^*E^*}$ such that $(d,e) \in (D^*, E^*)(\Omega) \implies \mathbb{P}_\alpha^{D^*E^*}(d,e) > 0$. Then by Theorem 2.28, $A \perp\!\!\!\perp_{\mathbb{P},ext} (B,D)|(C,E)$.

Only if:

For any β, $\mathbb{P}_\beta^{ABC|DE} = \mathbb{P}_\beta^{DE} \odot \mathbb{P}^{ABC|DE}$. By Lemma 2.24, we have $\overline{\mathbb{P}}^{A|BCDE}$ such that

$$\mathbb{P}_\beta^{A^*B^*C^*D^*E^*} = \mathbb{P}_\beta^{D^*E^*} \odot \mathbb{P}^{B^*C^*|D^*E^*} \odot \overline{\mathbb{P}}^{A^*|B^*C^*D^*E^*} \tag{124}$$

$$= \mathbb{P}_\beta^{B^*C^*D^*E^*} \odot \overline{\mathbb{P}}^{A^*|B^*C^*D^*E^*} \tag{125}$$

$$= \mathbb{P}_\beta^{C^*E^*} \odot \overline{\mathbb{P}}_\beta^{B^*D^*|C^*E^*} \odot \overline{\mathbb{P}}^{A^*|B^*C^*D^*E^*} \tag{126}$$

By Theorem 2.28, we have some α such that $\mathbb{P}_\alpha^{D^*E^*}$ is strictly positive on the range of (D^*, E^*) and $A^* \perp\!\!\!\perp_{\mathbb{P}_\alpha} (B^*, D^*)|(C^*, E^*)$.

By independence, for some version of $\overline{\mathbb{P}}^{A|BCDE}$:

$$(127)$$

$$(128)$$

$$= \overline{\mathbb{P}}_\alpha^{C^*E^*} \odot \overline{\mathbb{P}}_\alpha^{B^*D^*|C^*E^*} \odot (\overline{\mathbb{P}}_\alpha^{A^*|C^*E^*} \otimes \text{erase}_{BD}) \tag{129}$$

Thus for any $(a,b,c,d,e) \in A \times B \times C \times D \times E$ such that $\mathbb{P}_\alpha^{B^*C^*D^*E^*}(b,c,d,e) > 0$, $\overline{\mathbb{P}}^{A^*|B^*C^*D^*E^*}(a|b,c,d,e) = \overline{\mathbb{P}}_\alpha^{A^*|C^*E^*}(a|c,e)$ for any version of the disintegration.

However, by assumption, $\mathbb{P}_\alpha^{B^*C^*D^*E^*}(b,c,d,e) > 0 \implies \mathbb{P}_\beta^{B^*C^*D^*E^*}(b,c,d,e) > 0$, and so $\overline{\mathbb{P}}_\beta^{A^*|B^*C^*D^*E^*} \overset{a.s.}{=} disint P_\alpha^{A^*|C^*E^*}(a|c,e)$ also. But then

I actually haven't shown almost sure equality of disintegrations

$$\mathbb{P}_\beta^{A^*B^*C^*D^*E^*} = \text{(130)}$$

$$= \text{(131)}$$

$\square$

Definition 2.30 (Order 2 conditional independence). For a *probability 2-comb* $\mathbb{P}^{X|W \square Z|Y}$, define $\mathbb{P}_\gamma^{XYZ|W} := \text{insert}(\mathbb{Q}_\gamma^{Y|XW}, \mathbb{P}^{X|W \square Z|Y})$. Then say B is order 2 conditionally independent of A given C, written $B \perp\!\!\!\perp_\mathbb{P}^2 A|C$ if for all γ, $B \perp\!\!\!\perp_{\mathbb{P}_\gamma}^1 A|C$.

In each case, the semi-graphoid axioms hold for each lower level extension $\mathbb{P}_\alpha^I$, $\mathbb{P}_\beta^{VW}$ and $\mathbb{P}_\gamma^{XYZ|W}$, and so they also hold for the higher level definition of conditional probability.

2.9.4 Graphical properties of conditional independence

It is well-known that directed acyclic graphs are able to represent some conditional independence properties of probability models via the graphical property of *d-separation*. String diagrams are similar to directed acyclic graphs, and string diagrams can be translated into directed acyclic graphs and vise-versa (Fong, 2013). Thus we expect that a property analogous to d-separation can be defined for string diagrams.

We can reason from graphical properties of model disintegrations to graphical properties of models as Theorem 2.31. A general theory akin to d-separation for string diagrams may facilitate a more general understanding of how conditional independence properties of a model relate to conditional independence properties of its components.

Theorem 2.31. *Given a probability 2-comb* $\mathbb{P}^{\mathsf{XV}|\mathsf{W}\square\mathsf{Y}|\mathsf{XV}}$, $\mathsf{Y} \perp\!\!\!\perp^2_\mathbb{P} \mathsf{V}|\mathsf{WX}$ *if and only if there is a version of* $\overline{\mathbb{P}}^{\mathsf{Y}|\mathsf{WXV}}$ *and some* $\overline{\mathbb{P}}^{\mathsf{Y}|\mathsf{WX}}$ *such that*

$$\overline{\mathbb{P}}^{\mathsf{Y}|\mathsf{WXV}} = \begin{array}{c}\mathsf{W}\\\mathsf{X}\\\mathsf{V}\end{array} \boxed{\overline{\mathbb{P}}^{\mathsf{Y}|\mathsf{WX}}} \mathsf{Y} \tag{132}$$

Proof. If: For any $\mathbb{K}: X \times V \times W \twoheadrightarrow Y$, $\mathbb{J}: \{*\} \twoheadrightarrow W$

$$\mathbb{J} \odot \text{insert}(\mathbb{P}^{\mathsf{XV}|\mathsf{W}\square\mathsf{Y}|\mathsf{XV}}, \mathbb{K}) = \mathbb{J} \odot \mathbb{P}^{\mathsf{XV}|\mathsf{W}} \odot \mathbb{K} \odot \overline{\mathbb{P}}^{\mathsf{Y}|\mathsf{WXV}} \tag{133}$$
$$\tag{134}$$

But for any $\mathbb{L} := \mathbb{P}^{\mathsf{XV}|\mathsf{W}} \odot \mathbb{K}$ we have

Note that any $\mathbb{P}^{\mathsf{XV}|\mathsf{W}\square\mathsf{Y}|\mathsf{XV}}$ must, by assumption, be representable as $\qquad\square$

2.9.5 Restricted 2-combs

We're often interested in a subset of inserts to a 2-comb. We're interested in particular in inserts that depend on a subset of the "available" variables. There are in general many restrictions we could consider, but some we are more interested in than others. For a 2-comb restricted to a generic subset A of inserts, we will write $\mathbb{P}^{\mathsf{X}|\mathsf{W}\boxed{A}\mathsf{Y}|\mathsf{D}}$.

Given $\mathbb{P}^{\mathsf{X}|\mathsf{W}\square\mathsf{Y}|\mathsf{D}}$ and some random variable $\mathsf{V} = f \circ (\mathsf{W}, \mathsf{X})$, define the subset of inserts that depend only on V as those $\mathbb{Q}_\alpha^{\mathsf{D}|\mathsf{XW}}$ such that $\mathsf{D} \perp\!\!\!\perp^1_\mathbb{Q} (\mathsf{W}, \mathsf{X})|\mathsf{V}$. Define B as the set of all inserts exhibiting this conditional independence. Then write $\mathbb{P}^{\mathsf{X}|\mathsf{W}\boxed{\mathsf{V}}\mathsf{Y}|\mathsf{D}}$ for the restriction of $\mathbb{P}^{\mathsf{X}|\mathsf{W}\square\mathsf{Y}|\mathsf{D}}$ to this set. Note that B may be empty if there are no valid inserts with the required conditional independence. When we consider restrictions, we will assume that this set is non-empty.

A special case of some interest is the restriction $\mathbb{P}^{\mathsf{X}|\mathsf{W}\boxed{*}\mathsf{Y}|\mathsf{D}}$. For any (X, W), $* = \text{erase}_{X \times W} \circ (\mathsf{X}, \mathsf{W})$, so this is a well-defined restriction. However, we must still assume that there exists at lease one $\mathbb{Q}_\alpha^{\mathsf{D}|\mathsf{XW}}$ such that $\mathsf{D} \perp\!\!\!\perp^1_\mathbb{Q} (\mathsf{W}, \mathsf{X})|*$, which may not always be true (consider the example of height, weight and BMI above; there is no valid conditional probability in which BMI is independent of height and weight). This restriction is useful because conditional independences with respect to the restricted map can be interpreted as expressing the property that "unless we deliberately induce dependence, these variables are conditionally independent".

Definition 2.32 (Restricted order 2 conditional independence)**.** For a *restricted probability 2-comb* $\mathbb{P}^{\mathsf{X}|\mathsf{W}\boxed{E}\mathsf{Z}|\mathsf{Y}}$, define $\mathbb{P}_\gamma^{\mathsf{XYZ}|\mathsf{W}} := \text{insert}(\mathbb{Q}_\gamma^{\mathsf{Y}|\mathsf{XW}}, \mathbb{P}^{\mathsf{X}|\mathsf{W}\square\mathsf{Z}|\mathsf{Y}})$ for any

$\mathbb{Q}_\gamma \in E$. Then B is restricted order 2 conditionally independent of A given C, written $\mathsf{B} \perp\!\!\!\perp_{\mathbb{P}}^{2|E} \mathsf{A}|\mathsf{C}$ if for all $\gamma \in E$, $\mathsf{B} \perp\!\!\!\perp_{\mathbb{P}_\gamma}^{1} \mathsf{A}|\mathsf{C}$.

3 Decision theoretic causal inference

People very often have to make decisions with some information they may consult to help them make the decision. We are going to examine how gappy probability models can formally represent problems of this type, which in turn allows us to make use of the theory of probability to help guide us to a good decision. Probabilistic models have a long history of being used to represent decision problems, and there exist a number of coherence theorems that show that preferences that satisfy certain kinds of constraints must admit representation by a probability model and a utility function of the appropriate type. Particularly noteworthy are the theorems of Ramsey (2016) and Savage (1954), which together yield a method for representing decision problems known as "Savage decision theory", and the theorem of Bolker (1966); Jeffrey (1965) which yields a rather different method for representing decision problems known as "evidential decision theory". Joyce (1999) extends Jeffrey and Bolker's result to a representation theorem that subsumes both "causal decision theory" and "evidential decision theory".

None of these representation theorems explicitly concern themselves with probability gaps. One may try to find gappy probability models inside the theories, for example in the way that the Savage theory extends a probability distribution over *states* to a probability distribution over *consequence* when given an *act*. However, the connection to probability gaps is not explicit and may not follow precisely. We could make similar comments about causal or evidential decision theories: perhaps "causal conditionals" are gappy probability models, but perhaps they aren't exactly.

We do not have a comparable axiomatisation of preferences that yield a representation of decision problems in terms of utility and gappy probability. Such an undertaking could potentially clarify some choices that can be made in setting up a gappy probability model of decision making, but it is the subject of future work. Instead, we suppose that we are satisfied with a particular probabilistic model of a decision problem, a supposition based on convention rather than axiomatisation.

3.1 Decision problems

Suppose we have an observation process $\mathfrak{X}$, modelled by X taking values in X (we are *informed*). Given an observation $x \in X$, we suppose that we can choose a decision from a known set D (the set of decisions is *transparent*), and we suppose that choosing a decision results in some action being taken in the real world. As with processes of observation, we will mostly ignore the details of what "taking an action" involves. The process of choosing a decision that yields an element of D is a decision making process $\mathfrak{D}$ modelled by D. We might be

able to introduce randomness to the choice, in which case the relation between X and D may be stochastic. We will assume that there is some $\mathcal{Y}$ modelled by Y such that (X, D, Y) tell us everything we want to know for the purposes of deciding which outcomes are better than others.

We want a model that allows us to compare different stochastic *decision functions* $\mathbb{Q}_\alpha^{D|X} : X \rightarrow D$. What we mean by compare is: we need a higher-order function M that takes a decision function $\mathbb{Q}_\alpha^{X|D}$ and returns a probabilistic model of the consequences of selecting that decision function $\mathbb{P}_\alpha^{DXY} := M(\mathbb{Q}_\alpha^{D|X})$. A probability 2-comb $\mathbb{P}^{X \square Y|D}$ along with the "insert" operation previously defined is such a function, but there are many such functions that are not probability 2-combs. A general function M need not have, for example, $\mathbb{Q}_\alpha^{D|X} = \overline{\mathbb{P}}_\alpha^{D|X}$. We will define *ordinary decision problems* to be those for which the desired model M can be represented as probability 2-combs.

We allow for a further gap in our probability model; namely, we allow for the possibility that we don't know which probability 2-comb $\mathbb{P}^{X \square Y|D}$ we "ought" to use. We include an unobserved variable H, which stands for "hypothesis", which can be interpreted as a handle by which we can represent another gap in our knowledge: if we knew the value of H, or alternatively if we knew how H was distributed, then we would know that our decision problem was represented by a unique probability 2-comb $\mathbb{P}_h^{X \square Y|D}$. However, we do not assume a distribution over H is known, and so our model is given by $\mathbb{P}^{X|H \square Y|D}$.

Specifically, an *ordinary decision problem* M features hypotheses $H : \Omega \to H$, observations $X\Omega \to X$, decisions $D : \Omega \to D$ and consequences $Y : \Omega \to Y$. It is modeled by a function $M : \Delta(D)^X \to \Delta(X \times D \times Y)^H$ such that, defining $\mathbb{P}_\alpha^{DXY} := M(\mathbb{Q}_\alpha^{D|X})$,

1. There is some $\overline{\mathbb{P}}_\alpha^{D|X}$ such that $\mathbb{Q}_\alpha^{D|X} = \overline{\mathbb{P}}_\alpha^{D|X}$ for all decision functions α

2. There is some $\mathbb{P}^{X|H}$ such that, for all decision functions α, there is some $\overline{\mathbb{P}}_\alpha^{X|H}$ such that $\mathbb{P}^{X|H} = \overline{\mathbb{P}}_\alpha^{X|H}$

3. There is some $\mathbb{P}^{Y|XDH}$ such that, for all decision functions α, there is some $\overline{\mathbb{P}}_\alpha^{Y|XDH}$ such that $\mathbb{P}^{Y|XDH} = \overline{\mathbb{P}}_\alpha^{Y|XDH}$

(1) reflects the assumption that the "probability of D given X" based on the induced model is equal to the "probability of D given X" based on the chosen decision function. (2) reflects the assumption that the observations should be modelled identically no matter which decision function is chosen. (3) reflects the assumption that given hypothesis, the observations and the decision, the model of Y does not depend any further on the decision function α.

Under these assumptions, there exists a "see-do model" $\mathbb{P}^{X|H \square Y|D}$ where, for all α:

$$\mathbb{P}_\alpha = \quad \boxed{\mathbb{S}_\alpha^{D|X}} \quad \boxed{\mathbb{T}^{X|H}} \quad \boxed{\mathbb{T}^{Y|DHX}} \tag{135}$$

The proof is given in Appendix 6.

3.2 What should a probability model represent? Controversies about decision theories

As we have said, we do not prove any representation theorems showing that a decision problem must be representable as a probability 2-comb $\mathbb{P}^{X|H\square Y|D}$. Our motivations for this choice are convention and the fact that they seem intuitive for a number of causal problems. By convention, we mean that models of this form reduce with additional assumptions to well-known statistical and causal models. For an example of the intuitive appeal, we could consider for example a case where X represents medical trial data, D represents a recommendation to care providers and Y represents health indicators of interest. In such cases, it seems quite reasonable to assume that the decision rule (or inference rule) we choose to map data to recommendations is identical to the relationship between X and D that should appear in our model. It also seems reasonable to assume that our model should not consider the data collected to depend on the decision rule chosen, nor should it consider the consequences to have any additional dependence on the inference rule once the data and the decision have been fixed.

We could consider problems in which one or more of these assumptions could be argued to be unreasonable. Such questions are related to a number of controversies in decision theory. Newcomb's problem, for example, invites us to consider a problem where a second party has predicted our choice of decision function α before we have made the choice, and we have good reasons to believe this prediction is correct (Nozick, 1969). *The predictor* may then make choices that affect the consequences we expect to see, and this could mean that it is appropriate to consider models in which consequences to depend on α in addition to D. The question of which kind of model *should* be adopted in such a situation is controversial Weirich (2016); Lewis (1986), and two prominent views on the correct answer are *causal decision theory* and *evidential decision theory.*

This work does not propose normative rules for getting from a description of the world to a see-do model, and so the question of "which decision theory?" is not addressed here. We could ask if causal and evidential decision theory can be operationalised as different (vague) rules for getting from descriptions of the world to probability 2-combs $\mathbb{P}^{X|H\square Y|D}$, but we leave this question open.

31

3.3 Unresponsiveness

Given a see-do model $\mathbb{P}^{\mathsf{X|H}\square\mathsf{Y|D}}$, $\mathsf{A} \perp\!\!\!\perp_{\mathbb{P}}^{2|*} \mathsf{D|H}$ can be interpreted to express the property, given any hypothesis, A does not depend on D when we choose D randomly.

This might sound like it expresses a property of "causal independence". However, it isn't quite satisfactory for this term. Consider $\mathsf{A} \in \{0, 1\}$ representing the outcome of a fair coin toss, $\mathsf{D} \in \{0, 1\}$ representing a bet on the coin toss and $\mathsf{B} \in \{0, 1\}$ representing the outcome of the bet. There is one hypothesis – "the coin is fair" – and no observations. We construct a see-do model $\mathbb{P}$ in the obvious way given these assumptions. Then $\mathsf{A} \perp\!\!\!\perp_{\mathbb{P}}^{2|*} \mathsf{D|H}$ ("the coin toss is independent of the decision, for any decision rule") and $\mathsf{B} \perp\!\!\!\perp_{\mathbb{P}}^{2|*} \mathsf{D|H}$ ("the outcome of the bet is independent of the decision, for any decision rule"). However, it is not the case that $(\mathsf{A}, \mathsf{B}) \perp\!\!\!\perp_{\mathbb{P}}^{2|*} \mathsf{D|H}$. Were both A and B causally independent of D, then (A, B) also ought to be independent of D. This example is from Heckerman and Shachter (1995).

We will borrow the terminology of *unresponsiveness* from Heckerman and Shachter (1995) to refer to independences like $\mathsf{A} \perp\!\!\!\perp_{\mathbb{P}}^{2|*} \mathsf{D|H}$; specifically, where a variable is independent of D given H under the restricted 2-comb $\mathbb{P}^{\mathsf{X|H}\boxed{*}\mathsf{Y|D}}$.

3.4 Causal models similar to see-do models

Lattimore and Rohde (2019a) and Lattimore and Rohde (2019b) consider an observational probability model and a collection of indexed interventional probability models, with the probability model tied to the interventional models by shared parameters. In these papers, they show how such a model can reproduce inferences made using Causal Bayesian Networks. This kind of model can be identified with a type of see-do model, where what we call hypotheses H are identified with the sequence of what Rohde and Lattimore call parameter variables.

The approach to decision theoretic causal inference described by Dawid (2020) is somewhat different:

> 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.

The difference between the model described here and a see-do model is that a see-do model uses different variables X and Y to represent observations and consequences, while Dawid's model uses the same variable (T, Y) to represent outcomes in interventional and observational regimes. In this work we associate one observed variable with each measurement process, while in Dawid's approach (T, Y) seem to be doing double duty, representing mesurement processes

carried out during observations and after taking action. This can be thought of as the causal analogue of the difference betwen saying we have a sequence $(\mathsf{X}_1, \mathsf{X}_2, \mathsf{X}_3)$ of observations independent and identically distributed according to $\mu \in \Delta(X)$ and saying that we have some observations distributed according to $\mathbb{P}^\mathsf{X} \in \Delta(X)$. People usually understand what is meant by the latter, but if one is trying to be careful the former is a more precise statement of the model in question.

We have already noted a connection between our work and Heckerman and Shachter (1995). Our approach is quite close to their approach if we identify what we call hypotheses with what they call states and allow for probabilistic dependence between states, decisions and consequences. It is an open question whether their notion of limited unresponsiveness corresponds to one of our notions of conditional independence.

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

3.5 See-do models and classical statistics

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

A further property that may hold for some see-do models $\mathbb{P}^{\mathsf{X}|\mathsf{H}\square\mathsf{Y}|\mathsf{D}}$ is $\mathsf{Y} \perp\!\!\!\perp_\mathbb{P}^2 \mathsf{X}|(\mathsf{H},\mathsf{D})$. This expresses the view that the consequences are independent of the observations, once the hypothesis and the decision are fixed. Such a situation could hold in our scenario above, where the observations are trial data, the decisions are recommendations to care providers and the consequences are future patient outcomes. In such a situation, we might suppose that the trial data are informative about the consequences only via some parameter such as effect size; if the effect size can be deduced from H then our assumption corresponds to the conditional independence above.

Given a see-do model $\mathbb{P}^{\mathsf{X}|\mathsf{H}\square\mathsf{Y}|\mathsf{D}}$ along with the principle of expected utility to evaluate strategies, and the assumption $\mathsf{Y} \perp\!\!\!\perp_\mathbb{P}^2 \mathsf{X}|(\mathsf{H},\mathsf{D})$ we obtain a statistical decision problem in the form introduced by Wald (1950).

A *statistical model* (or *statistical experiment*) is a collection of probability distributions $\{\mathbb{P}_\theta\}$ indexed by some set Θ. A statistical decision problem gives us an observation variable $\mathsf{X} : \Omega \twoheadrightarrow X$ and a statistical experiment $\{\mathbb{P}_\theta^\mathsf{X}\}_\Theta$, a decision set D and a loss $l : \Theta \times D \to \mathbb{R}$. A strategy $\mathbb{S}_\alpha^{\mathsf{D}|\mathsf{X}}$ is evaluated according

to the risk functional $R(\theta, \alpha) := \sum_{x \in X} \sum_{d \in D} \mathbb{P}_\theta^{\mathsf{X}}(x) S_\alpha^{\mathsf{D}|\mathsf{X}}(d|x) l(h,d)$. A strategy $\mathbb{S}_\alpha^{\mathsf{D}|\mathsf{X}}$ is considered more desirable than $\mathbb{S}_\beta^{\mathsf{D}|\mathsf{X}}$ if $R(\theta, \alpha) < R(\theta, \beta)$.

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

$$\mathbb{E}_{\mathbb{P}_\alpha}[\mathsf{Y}|h] = \sum_{x \in X} \sum_{d \in D} \sum_{y \in Y} \mathbb{P}^{\mathsf{X}|\mathsf{H}}(x|h) \mathbb{Q}_\alpha^{\mathsf{D}|\mathsf{X}}(d|x) \overline{\mathbb{P}}^{\mathsf{Y}|\mathsf{HD}}(y|h,d) \tag{136}$$

$$= \sum_{x \in X} \sum_{d \in D} \mathbb{P}^{\mathsf{X}|\mathsf{H}}(x|h) \mathbb{Q}_\alpha^{\mathsf{D}|\mathsf{X}}(d|x) l(h,d) \tag{137}$$

$$= R(h, \alpha) \tag{138}$$

If we are given a see-do model where we interpret $\{\mathbb{P}^{\mathsf{X}|\mathsf{H}}(\cdot|h)|h \in H\}$ as a statistical experiment and Y as a negative utility, the expectation of the utility under the strategy forecast given in equation 135 is the risk of that strategy under hypothesis h.

4 Causal Bayesian Networks

Like some of the causal modelling frameworks discussed in the previous section, including see-do models, Causal Bayesian Networks (CBNs) represent both "observations" and "consequences of interventions". It seems reasonable to think that the real-world things that the see-do framework and the CBN framework address are sometimes the same. The question we have here is: if we have a decision problem represented by a see-do model, when can we represent the same problem with a CBN?

In order to answer this question, we have to deal with the fact that neither theory is formally contained by the other, so there's no precise way, for example, for decisions to correspond to interventions. The correspondence exists in the territory, the world that is inhabited by measurement processes, not the precise mathematical world that is inhabited by random variables. We therefore have to make some choices about what corresponds to what that seem to be reasonable given our understanding of what these models are used for.

First, we will consider what kind of model is represented by a CBN; we will argue that they are in general probability models with order 2 gaps. We will assume a sample space Ω and a collection of variables, and we will consider a simplified kind of CBN where a single variable may be intervened on. With this condition, according to Pearl (2009), a causal Bayesian network is a probability model $\mathbb{P}$, a collection of interventional probability models $\{\mathbb{P}_{\mathsf{X}=a}|a \in X_i\}$ and a directed acyclic graph $\mathcal{G}$ whose nodes are identified with variables. Given one intervenable variable, we can in general consider just three variables $\{\mathsf{W}, \mathsf{X}, \mathsf{Y}\}$, where W is all of the ancestors of X in $\mathcal{G}$, X is the "intervenable" variable and Y are all the other variables associated with nodes of $\mathcal{G}$. The interventional

probability models must all obey the truncated factorisation condition with respect to $\mathcal{G}$:

$$\mathbb{P}^{\mathsf{WXY}}_{\mathsf{X}\bowtie a}(w,x,y) = \mathbb{P}^{\mathsf{W}}(w)\mathbb{P}^{\mathsf{Y}|\mathsf{XW}}(y|x,w)[\![x=a]\!] \tag{139}$$

We will interpret observations and interventions as follows: we have a sequence of "observation" variables $\mathsf{V}_{[n]} := (\mathsf{W}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i \in [n]}$ mutually independent and identically distributed and a sequence of "consequence" variables modeled by $\mathsf{V}_{[n,m]} := (\mathsf{W}_i, \mathsf{X}_i, \mathsf{Y}_i)_{i \in [n,m]}$ mutually independent and identically distributed for any choice of intervention $\mathsf{X} \bowtie a$, and also independent of $\mathsf{V}_{[n]}$ given a choice of intervention. If we were given the fact that the observations were distributed according to $\mathbb{P}^{\mathsf{WXY}} \in \Delta(W \times X \times Y)$, then for any $\mathsf{X} \bowtie a$ the consequences will each be distributed according to $\mathbb{P}^{\mathsf{WXY}}_{\mathsf{X}\bowtie a} \in \Delta(W \times X \times Y)$, related to $\mathbb{P}^{\mathsf{WXY}}$ via Equation 139.

If we want some model $\mathbb{T}$ of the variables $\mathsf{V}_{[m]}$, we therefore have for $i \in [n]$, $j \in [m,n]$:

$$\mathbb{T}^{\mathsf{X}_i\mathsf{W}_i\mathsf{Y}_i} = \mathbb{P}^{\mathsf{WXY}} \tag{140}$$

$$\mathbb{T}^{\mathsf{W}_j\mathsf{X}_j\mathsf{Y}_j|\mathsf{X}_j}(w,x,y|a) = \mathbb{P}^{\mathsf{WXY}}_{\mathsf{X}\bowtie a}(w,x,y) \tag{141}$$

$$\implies \overline{\mathbb{T}}^{\mathsf{W}_j\mathsf{X}_j\mathsf{Y}_j|\mathsf{X}_j} = \mathbb{T}^{\mathsf{W}_i}(w)\mathbb{T}^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{W}_i}(y|x,w)[\![x=x]\!] \tag{142}$$

We assumed independence of V_js from $\mathsf{V}_{[n]}$ for $j \in [n,m]$; that is $\mathsf{V}_{[n,m]} \perp\!\!\!\perp^2_{\mathbb{T}}$ $[V]_{[n]}|\mathsf{X}_{[n,m]}$.

$$\mathbb{T}^{\mathsf{W}_j\mathsf{X}_j\mathsf{Y}_j|\mathsf{X}_j} \tag{143}$$

These equations tell us that if we were given $\mathbb{P}^{\mathsf{WXY}}$ by an oracle, then we would have a proability 2-comb $\mathbb{T}^{\mathsf{V}_{[n]}\square\mathsf{V}_{[n,m]}|\mathsf{X}_{[n,m]}}$. However, "if we were given $\mathbb{P}^{\mathsf{WXY}}$ by an oracle" is much the same as saying "$\mathbb{P}^{\mathsf{WXY}}$ is a probability gap".

The proposition "there is an oracle that tells us $\mathbb{P}^{\mathsf{WXY}}$" sounds a bit like we have a measurment process for determining $\mathbb{P}^{\mathsf{WXY}}$ – namely, consult the oracle. However this is not an actual process; instead, the injunction to consult the oracle is actually saying *if we had a measurement process for $\mathbb{P}^{\mathsf{WXY}}$*, then here is how we would proceed. In the previous section, we represented a gap in our knowledge of this type with a hypothesis variable H along with a model $\mathbb{T}$ such that $\mathbb{T}^{\mathsf{H}}$ is a probability gap, and what is referred to by the CBN setup as $\mathbb{P}^{\mathsf{WXY}}$ corresponds to $\mathbb{T}^{\mathsf{W}_i\mathsf{X}_i\mathsf{Y}_i|\mathsf{H}}(\cdot|h)$ for $i \in [n]$ – that is, the probability distribution that we would use for $(\mathsf{W}_i, \mathsf{X}_i, \mathsf{Y}_i)$ if the oracle told us that $\mathsf{H} \bowtie h$.

We have variables $\mathsf{V}_{[n]}$, $\mathsf{V}_{[n,m]}$ and H with some model $\mathbb{T}$ that has a probability gap for $\mathbb{T}^{\mathsf{H}}$. What kind of model is $\mathbb{T}$? We think it is appropriate to consider that it has an additional probability gap. Equation 139 allows us to define $\mathbb{T}^{\mathsf{V}_{[n,m]}|\mathsf{HX}_{[n,m]}}$ from $\mathbb{T}^{\mathsf{V}_{[n]}|\mathsf{H}}$. In particular, for any $i \in [n]$, $j \in [n,m]$

$$\mathbb{T}^{\mathsf{V}_j|\mathsf{HX}_{[n,m]}}(w,x',y|h,x) = \mathbb{T}^{\mathsf{W}_i|\mathsf{H}}(w|h)\mathbb{T}^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{W}_i}(y|x,w)[\![x=x]\!] \tag{144}$$

$\mathbb{T}^{V_{[n,m]}|HX_{[n,m]}}$ and $\mathbb{T}^{V_{[n]}|H}$ together define a probability 2-comb $\mathbb{T}^{V_{[n]}|H\square V_{[n,m]}|X_{[n,m]}}$. The family of inserts for this 2-comb are conditional probabilities of the form $\mathbb{U}_{\alpha}^{X_{[n,m]}|HV_{[n]}W_{[n,m]}}$.

Decision rules do not necessarily take the form $\mathbb{U}_{\alpha}^{X_{[n,m]}|HV_{[n]}W_{[n,m]}}$.

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

$$\mathbb{P}_{\alpha}^{V_B|H} :=$$

$$(145)$$

$$= $$

$$(146)$$

$$=: \mathbb{Q}_{\alpha}^{V_B|H} \qquad (147)$$

Theorem 4.1. *Given a standard see-do model* $\mathcal{M} := \{\mathbb{T}^{OV_B|HD}, \{\mathbb{S}_{\alpha}^{D|V_A}|\alpha \in \mathcal{A}\}\}$ *and a CBN model* $\mathbb{U}^{V_AV_j|HX_j}$ *as defined in Equation* **??**, *then there exists a see-do model* $\{\mathbb{U}^{V_iV_j|HX_j}, \{\mathbb{R}_{\alpha}^{X_j|V_AW_jH}|\alpha \in \mathcal{A}\}\}$ *consistent with* $\mathcal{M}$ *if and only if*

1. W_j *is a pre-choice variable, i.e.* $(V_A, W_j) \perp\!\!\!\perp_{\mathbb{T}} D|H$

2. $\mathbb{T}^{V_AW_j|H} = \mathbb{U}^{V_AW_j|H}$

3. $Y_j \perp\!\!\!\perp_{\mathbb{T}} D|W_jV_AHX_j$

4. $\mathbb{T}^{Y_j|W_jV_AHX_j} = \mathbb{U}^{Y_j|W_jV_AHX_j}$

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

$$\mathbb{T}^{V_AV_j|HD} =$$

$$(148)$$

For each $\mathbb{S}_{\alpha}^{D|V_A}$, define

$$\mathbb{R}_{\alpha}^{X_j|V_AW_jH} :=$$

$$(149)$$

36

Then

$$\text{(150)}$$

$$\text{(151)}$$

$$\text{(152)}$$

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

$$\mathbb{P}_d^{\mathsf{W}_j|\mathsf{H}}(w|h) = \mathbb{T}^{\mathsf{W}_j|\mathsf{HD}}(w_j|h,d) \tag{153}$$

$$\neq \mathbb{T}^{\mathsf{W}_j|\mathsf{HD}}(w_j|h,d') \tag{154}$$

$$= \mathbb{P}_{d'}^{\mathsf{W}_j|\mathsf{H}}(w|h) \tag{155}$$

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

$$\mathbb{P}_\alpha^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}} = \mathbb{T}^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}} \tag{156}$$

$$\neq \mathbb{U}^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}} \tag{157}$$

$$= \mathbb{Q}_\alpha^{\mathsf{V}_A\mathsf{W}_j|\mathsf{H}} \tag{158}$$

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

$$\mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{HX}_j\mathsf{D}}(y|w,v,h,x,d) \neq \mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{HX}_j\mathsf{D}}(y|w,v,h,x,d') \tag{159}$$

$$\text{and } \mathbb{T}^{\mathsf{X}_j\mathsf{W}_j\mathsf{V}_A|\mathsf{HD}}(x,w,v|h,d) > 0 \tag{160}$$

$$\text{and } \mathbb{T}^{\mathsf{X}_j\mathsf{W}_j\mathsf{V}_A|\mathsf{HD}}(x,w,v|h,d') > 0 \tag{161}$$

$$\tag{162}$$

The latter conditions hold as if Equation 159 only held on sets of measure 0 then we could choose versions of the conditional probabilities such that the independence held.

Then

$$\mathbb{P}_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) = \mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x,d) \tag{163}$$

$$\neq \mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x,d') \tag{164}$$

$$= \mathbb{P}_{d'}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) \tag{165}$$

$$\implies \mathbb{P}_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) \neq \mathbb{Q}_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) \tag{166}$$

$$\text{or } \mathbb{P}_{d'}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) \neq \mathbb{Q}_{d'}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) \tag{167}$$

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

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

$$\mathbb{P}_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) = \mathbb{T}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j}(y|w,v,h,x) \tag{168}$$

$$\neq \mathbb{U}^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j}(y|w,v,h,x) \tag{169}$$

$$\neq modelQ_d^{\mathsf{Y}_j|\mathsf{W}_j\mathsf{V}_A\mathsf{H}\mathsf{X}_j\mathsf{D}}(y|w,v,h,x) \tag{170}$$

$$\square$$

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

4.1 Proxy control

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

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

$$\mathbb{T}^{\mathsf{V}_A\mathsf{V}_B|\mathsf{HD}} = \qquad\qquad\qquad\qquad\qquad\qquad\qquad (171)$$

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

$$(172)$$

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

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

5 Potential outcomes

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

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

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

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

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

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

Proof. Define $\mathsf{Y}^X := (\mathsf{Y})_{x \in X}$. Then we can consistently extend $\mathbb{K}^{\mathsf{YX}|\mathsf{W}}$ to $\mathbb{K}^{\mathsf{YXY}^X|\mathsf{W}}$ by repeated application of Lemma 2.22. $\qquad\square$

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

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

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

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

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

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

$$\mathbb{P}^{\mathsf{XYY}^X|\mathsf{H}} = \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (173)$$
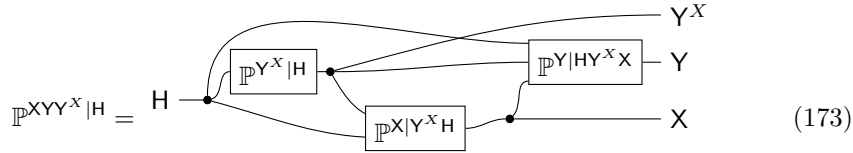
For any collection of representative kernels $\mathbb{T}^{\mathsf{Y}^X|\mathsf{H}}$, $\mathbb{T}^{\mathsf{X}|\mathsf{Y}^X\mathsf{H}}$ and $\mathbb{T}^{\mathsf{Y}|\mathsf{HY}^X\mathsf{X}}$. We can simplify Equation 173 somewhat. Firstly, $\mathbb{P}^{\mathsf{Y}|\mathsf{HY}^X\mathsf{X}}$ must always be represented a *selector kernel* $\varPi : X \times Y^{|X|} \rightarrow Y$, as shown by Lemma 5.5.

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

$$\mathbb{Q} := \quad \begin{array}{c} Y^X \\ X \\ W \end{array} \!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\! \boxed{\varPi} \!\!\!\! - Y \qquad\qquad\qquad\qquad (174)$$

$$\Longleftrightarrow \qquad\qquad\qquad\qquad\qquad\qquad (175)$$

$$\mathbb{Q}^y_{(y^X,x,w)} = \varPi^y_{(x,y^X)} \qquad\qquad\qquad \forall y, y^X, x, w \qquad (176)$$

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

Proof. Recall $\mathsf{Y} = \pi \circ (\mathsf{X}, \mathsf{Y}^X)$. Thus consistency implies that $\mathsf{Y} \overset{a.s.}{=} \pi \circ (\mathsf{X}, \mathsf{Y}^X)$ for all $(x, y^X, w) \in \mathrm{Range}(\mathsf{X}) \times \mathrm{Range}(\mathsf{Y}) \times \mathrm{Range}(\mathsf{W})$ such that $\mathsf{X}^{-1}(x) \cap (\mathsf{Y}^X)^{-1}(y^X) \cap \mathsf{W}^{-1}(w) \neq \emptyset$. However, wherever $\mathsf{X}^{-1}(x) \cap (\mathsf{Y}^X)^{-1}(y^X) \cap \mathsf{W}^{-1}(w) = \emptyset$, consistency implies $\mathbb{T}^{\mathsf{YY}^X\mathsf{X}|\mathsf{W}}(y, y^X, x|w) = 0$ and so $\mathbb{T}^{\mathsf{Y}|\mathsf{Y}^X\mathsf{XW}}$ is arbitrary on this collection of values. Equations 174 and 176 are equivalent to the statement $\mathsf{Y} \overset{a.s.}{=} \pi \circ (\mathsf{X}, \mathsf{Y}^X)$. $\qquad\square$

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

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

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

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

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

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

$$\mathbb{T}^{\mathsf{Y}|\mathsf{H}\mathsf{W}\mathsf{X}\mathsf{D}}(y|h, w, 0, 0) \neq \mathbb{T}^{\mathsf{Y}|\mathsf{H}\mathsf{W}\mathsf{X}\mathsf{D}}(y|h, w, 0, 2) \tag{177}$$

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

$$\mathbb{P}_0^{\mathsf{Y}|\mathsf{H}\mathsf{D}}(y|h, 0) = \sum_{x \in \{0,1\}, w \in W} \mathbb{T}^{\mathsf{W}|\mathsf{H}}(w|h) \mathbb{S}_0^{\mathsf{D}|\mathsf{W}}(0|w) \mathbb{T}^{\mathsf{X}|\mathsf{H}\mathsf{W}\mathsf{D}}(x|h, w, 0) \mathbb{T}^{\mathsf{Y}|\mathsf{H}\mathsf{W}\mathsf{X}\mathsf{D}}(y|h, w, x, 0)$$
$$\tag{178}$$

$$= \mathbb{T}^{\mathsf{Y}|\mathsf{H}\mathsf{W}\mathsf{X}\mathsf{D}}(y|h, w, 0, 0) \tag{179}$$

$$\neq \mathbb{P}_2^{\mathsf{Y}|\mathsf{H}\mathsf{D}} \tag{180}$$

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

assumption

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

$$\tag{181}$$

$$= \sum_{y^X} \mathbb{T}^{\mathsf{Y}^X|\mathsf{H}}(y^X|h)\Pi(y|0,y^X) \tag{182}$$

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

$$\tag{183}$$

$$= \mathbb{P}_2^{\mathsf{Y}|\mathsf{HD}} \tag{184}$$

Here we use the property $\mathsf{Y}^X \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$, implied by the assumption that Y^X is a prechoice variable. Equations 180 and 184 are clearly contradictory, thus there can be no potential outcomes Y^X in this example.

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

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

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

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

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

$$\mathsf{Y} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{WX} \tag{185}$$

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

Note the similarity between Equation 185 and the condition for proxy control in the previous section. Indeed, the two are identical if we identify U with $(\mathsf{W}_j, \mathsf{V}_A, \mathsf{X}_j)$.

6 Appendix:see-do model representation

> Update notation

Theorem 6.1 (See-do model representation). *Suppose we have a decision problem that provides us with an observation $x \in X$, and in response to this we can select any decision or stochastic mixture of decisions from a set D; that is we can choose a "strategy" as any Markov kernel $\mathbb{S} : X \to \Delta(D)$. 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 $\mathbb{S}$ as a joint probability over observations, decisions and consequences $\mathbb{P}_{h,\mathbb{S}} \in \Delta(X \times D \times Y)$.*

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: $\mathbb{P}_{h,\mathbb{S}}[\mathsf{X}] = \mathbb{P}_{h,\mathbb{S}''}[\mathsf{X}]$ for all $h, \mathbb{S}, \mathbb{S}'$ (observations are given "before" our strategy has any effect)*

2. *The chosen strategy is a version of the conditional probability of decisions given observations: $\mathbb{S} = \mathbb{P}_{h,\mathbb{S}}[\mathsf{D}|\mathsf{X}]$*

3. *There exists some strategy $\mathbb{S}$ that is strictly positive*

4. *For any $h \in H$ and any two strategies $\mathbb{Q}$ and $\mathbb{S}$, we can find versions of each disintegration such that $\mathbb{P}_{h,\mathbb{Q}}[\mathsf{Y}|\mathsf{DX}] = \mathbb{P}_{h,\mathbb{S}}[\mathsf{Y}|\mathsf{DX}]$ (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 $(\mathbb{T}, \mathsf{H}', \mathsf{D}', \mathsf{X}', \mathsf{Y}')$ such that $\mathbb{P}_{h,\mathbb{S}}[\mathsf{XDY}]^{ijk} = \mathbb{T}[\mathsf{X}'|\mathsf{H}']_h^i \mathbb{S}_i^j \mathbb{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']_{ijk}^k$.

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

$$\mathbb{P}[\mathsf{XDY}]^{ijk} = \mathbb{P}[\mathsf{X}]^i \mathbb{P}[\mathsf{D}|\mathsf{X}]_i^j \mathbb{P}[\mathsf{Y}|\mathsf{XD}]_{ij}^k \tag{186}$$

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

$$\mathbb{T}_{hj}^{kl} = \mathbb{P}_{h,\mathbb{S}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbb{S}}[\mathsf{Y}|\mathsf{XD}]_{kj}^l \tag{187}$$

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

$$\mathbb{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj} = \mathbb{T}\underline{\mathsf{X}'}^k_{hj} \tag{188}$$

$$= \sum_l \mathbb{T}^{kl}_{hj} \tag{189}$$

$$= \mathbb{P}_{h,\mathbb{S}}[\mathsf{X}]^k \tag{190}$$

$$= \mathbb{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj'} \tag{191}$$

Thus $\mathsf{X}' \perp\!\!\!\perp_\mathbb{T} \mathsf{D}'|\mathsf{H}'$ and so $\mathbb{T}[\mathsf{X}'|\mathsf{H}']$ exists (section 2.9.3) and $(\mathbb{T}, \mathsf{H}', \mathsf{D}', \mathsf{X}', \mathsf{Y}')$ is a see-do model.

Applying Equation 186 to $\mathbb{P}_{h,\mathbb{S}}$:

$$\mathbb{P}_{h,\mathbb{S}}[\mathsf{XDY}]^{ijk} = \mathbb{P}_{h,\mathbb{S}}[\mathsf{X}]^i \mathbb{P}_{h,\mathbb{S}}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{P}_{h,\mathbb{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \tag{192}$$

$$= \mathbb{P}_{h,\mathbb{S}}[\mathsf{X}]^i \mathbb{P}_{h,\mathbb{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \tag{193}$$

$$= \mathbb{P}_{h,\mathbb{S}}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{T}[\mathsf{X}'\mathsf{Y}'|\mathsf{H}'\mathsf{D}']^{ik}_{hj} \tag{194}$$

$$= \mathbb{S}^j_i \mathbb{T}[\mathsf{X}'\mathsf{Y}'|\mathsf{H}'\mathsf{D}']^{ik}_{hj} \tag{195}$$

$$= \mathbb{S}^j_i \mathbb{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^i_{hj} \mathbb{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} \tag{196}$$

$$= \mathbb{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbb{S}^j_i \mathbb{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} \tag{197}$$

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

$$\mathbb{P}_{h,\mathbb{S}}[\mathsf{X}]^i = \mathbb{P}_{h,\mathbb{Q}}[\mathsf{X}]^i \tag{198}$$

$$\mathbb{P}_{h,\mathbb{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} = \mathbb{P}_{h,\mathbb{Q}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \text{ for some version of } \mathbb{P}_{h,\mathbb{Q}}[\mathsf{Y}|\mathsf{XD}] \tag{199}$$

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

$$\mathbb{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbb{Q}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbb{Q}}[\mathsf{Y}|\mathsf{XD}]^l_{kj} \tag{200}$$

Then by substitution of $\mathbb{Q}$ for $\mathbb{S}$ in Equation 192 and working through the same steps

$$\mathbb{P}_{h,\mathbb{S}}[\mathsf{XDY}]^{ijk} = \mathbb{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbb{Q}^j_i \mathbb{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} \tag{201}$$

As $\mathbb{Q}$ was arbitrary, this holds for all strategies. $\qquad\square$

7 Appendix: Counterfactual representation

Definition 7.1 (Parallel potential outcomes)**.** Given a Markov kernel space $(\mathbb{K}, E, F)$, 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 $\mathbb{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]$ exists and $\mathbb{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]_{kj_1 j_2 \ldots j_{|W|}} = \delta[j_k]$.

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

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

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

7.1 Parallel potential outcomes representation theorem

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

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

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

The representation theorem for potential outcomes is somewhat simpler that

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

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

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

Definition 7.2 (Exchangeability). Given a Markov kernel space $(\mathbb{K}, E, F)$, 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 $\mathsf{D}_{[n]} = (\mathsf{D}_i)_{i \in [n]}$, $\mathbb{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$ exists and for any bijection $\pi : [n] \to [n]$ $\mathbb{K}[\mathsf{Y}_{\pi([n])}|\mathsf{D}_{\pi([n])}\mathsf{Z}] = \mathbb{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$.

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

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

Definition 7.4 (Extendably exchangeable). Given a Markov kernel space $(\mathbb{K}, E, F)$, 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 $\mathbb{K}$ 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 $(\mathbb{K}, E, F)$ while $((\mathsf{D}_i, \mathsf{Y}_i))_{i \in \mathbb{N} \setminus [n]}$ may be defined only on the extension.

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

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

Definition 7.5 (Deterministically reproducible)**.** Given a Markov kernel space $(\mathbb{K}, E, F)$, 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 $n \geq 2$, $\mathbb{K}[\mathsf{Y}_{[n]} | \mathsf{D}_{[n]} \mathsf{Z}]$ exists and $\mathbb{K}[\mathsf{Y}_{\{i,j\}} | \mathsf{D}_{\{i,j\}} \mathsf{Z}]_{kk}^{lm} = [\![l = m]\!] \mathbb{K}[\mathsf{Y}_i | \mathsf{D}_i \mathsf{Z}]_k^l$ for all i, j, k, l, m.

Theorem 7.6 (Potential outcomes representation)**.** *Given a Markov kernel space $(\mathbb{K}, E, F)$ along with a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i \in [n]}$ with $n \geq 2$ and a conditioning variable Z, $(\mathbb{K}, E, F)$ 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 $\mathbb{K}[\mathsf{Y}(D) | \mathsf{D}_{[D]} \mathsf{Z}]_{ez} = \mathbb{K}[\mathsf{Y}_D | \mathsf{D}_D \mathsf{Z}]_{ez}$ and introduce X_i, $i \in D$ such that $\mathbb{K}[\mathsf{X}_i | \mathsf{D}_i \mathsf{Z} \mathsf{Y}(D)]_{e_i z j_1 \ldots j_{|D|}}^{x_i} = \delta[j_{e_i}]^{x_i}$. Clearly $\{\mathsf{X}_{[n]}, \mathsf{D}_{[n]}, \mathsf{Y}(D)\}$ is a parallel potential outcome submodel. We aim to show that $\mathbb{K}[\mathsf{Y}_{[n]} | \mathsf{D}_{[n]} \mathsf{Z}] = \mathbb{K}[\mathsf{X}_{[n]} | \mathsf{D}_{[n]} \mathsf{Z}]$.

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

$$\mathbb{K}[\mathsf{X}_n | \mathsf{D}_n \mathsf{Z}]_{dz}^x = \sum_{y \in Y^{|D|}} \mathbb{K}[\mathsf{X}_{[n]} | \mathsf{D}_n \mathsf{Z} \mathsf{Y}(D)]_{dzy}^x \mathbb{K}[\mathsf{Y}(D) | \mathsf{D}_{[n]} \mathsf{Z}]_{dz}^y \tag{202}$$

$$= \sum_{y \in Y^{|D|}} \prod_{i \in [n]} \delta[y_{d_i}]^{x_i} \mathbb{K}[\mathsf{Y}(D) | \mathsf{D}_n \mathsf{Z}]_{dz}^y \tag{203}$$

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 $\mathbb{K}[\mathsf{Y}_{[n]} | \mathsf{D}_{[n]} \mathsf{Z}]_d z^x = 0$ 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 $\mathsf{Y}(S_d) = (\mathsf{Y}_d)_{d \in S_d}$.

$$\mathbb{K}[\mathsf{X}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]_{dz}^x = \sum_{y \in Y^{|D|}} \prod_{i \in Q_d} \delta[y_{d_i}]^{x_i} \mathbb{K}[\mathsf{Y}(D)|\mathsf{D}_{[n]}\mathsf{Z}]_{dz}^y \tag{204}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbb{K}[\mathsf{Y}(R_d)\mathsf{Y}(R_d^C)|\mathsf{D}_{Q_d}\mathsf{D}_{Q_d^C}\mathsf{Z}]_{d_{Q_d} d_{Q_d}^C z}^{x_{Q_d} y'} \tag{205}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbb{K}[\mathsf{Y}_{R_d}\mathsf{Y}_{R_d^C}|\mathsf{D}_{Q_d}\mathsf{D}_{Q_d^C}\mathsf{Z}]_{dz}^{x_{Q_d} y'} \tag{206}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbb{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]_{dz}^{x_{Q_d} y'} \qquad \text{(using exchangeability)} \tag{207}$$

Note that

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

8 Appendix: Connection is associative

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

Define

$$\mathsf{I}_{K\cdot\cdot} := \mathsf{I}_K \setminus \mathsf{I}_L \setminus \mathsf{I}_J \tag{208}$$

$$\mathsf{I}_{KL\cdot} := \mathsf{I}_K \cap \mathsf{I}_L \setminus \mathsf{I}_J \tag{209}$$

$$\mathsf{I}_{K\cdot J} := \mathsf{I}_K \cap \mathsf{I}_J \setminus \mathsf{I}_L \tag{210}$$

$$\mathsf{I}_{KLJ} := \mathsf{I}_K \cap \mathsf{I}_L \cap \mathsf{I}_J \tag{211}$$

$$\mathsf{I}_{\cdot L\cdot} := \mathsf{I}_L \setminus \mathsf{I}_K \setminus \mathsf{I}_J \tag{212}$$

$$\mathsf{I}_{\cdot LJ} := \mathsf{I}_L \cap \mathsf{I}_J \setminus \mathsf{I}_K \tag{213}$$

$$\mathsf{I}_{\cdot\cdot J} := \mathsf{I}_J \setminus \mathsf{I}_K \setminus \mathsf{I}_L \tag{214}$$

$$\mathsf{O}_{K\cdot\cdot} := \mathsf{O}_K \setminus \mathsf{I}_N \setminus \mathsf{I}_J \tag{215}$$

$$\mathsf{O}_{KL\cdot} := \mathsf{O}_K \cap \mathsf{I}_L \setminus \mathsf{I}_J \tag{216}$$

$$\mathsf{O}_{K\cdot J} := \mathsf{O}_K \cap \mathsf{I}_J \setminus \mathsf{I}_L \tag{217}$$

$$\mathsf{O}_{KLJ} := \mathsf{O}_K \cap \mathsf{I}_L \cap \mathsf{I}_J \tag{218}$$

$$\mathsf{O}_{L\cdot} := \mathsf{O}_L \setminus \mathsf{I}_J \tag{219}$$

$$\mathsf{O}_{LJ} := \mathsf{O}_L \cap \mathsf{I}_J \tag{220}$$

Also define

$$(\mathbb{P}, \mathsf{I}_P, \mathsf{O}_P) := \mathbb{K} \rightrightarrows \mathbb{L} \tag{221}$$

$$(\mathbb{Q}, \mathsf{I}_Q, \mathsf{O}_Q) := \mathbb{L} \rightrightarrows \mathbb{J} \tag{222}$$

Then

$$(\mathbb{K} \rightrightarrows \mathbb{L}) \rightrightarrows \mathbb{J} = \mathbb{P} \rightrightarrows \mathbb{J} \tag{223}$$

$$\tag{224}$$

$$\tag{225}$$

$$\tag{226}$$

$$\tag{227}$$

$$= \mathbb{K} \rightrightarrows (\mathbb{L} \rightrightarrows \mathbb{J}) \tag{228}$$

9   Appendix: String Diagram Examples

Recall the definition of *connection*:

Definition 9.1 (Connection)**.**

$$\mathbb{K} \rightrightarrows \mathbb{L} := \quad \begin{matrix} \mathsf{I}_{F.} \\ \mathsf{I}_{FS} \\ \mathsf{I}_{.S} \end{matrix} \boxed{\mathbb{K}} \begin{matrix} \mathsf{O}_{F.} \\ \mathsf{O}_{FS} \\ \boxed{\mathbb{L}} \ \mathsf{O}_S \end{matrix} \tag{229}$$

$$:= \mathbb{J} \tag{230}$$

$$\mathbb{J}^{zxw}_{yqr} = \mathbb{K}^{zx}_{yq}\mathbb{L}^{w}_{xqr} \tag{231}$$

Equation 229 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:

$$(\mathbb{J}, (\mathsf{I}_{F.}, \mathsf{I}_{FS}, \mathsf{I}_{.S}), (\mathsf{O}_{F.}, \mathsf{O}_{FS}, \mathsf{O}_S)) = \begin{bmatrix} \mathsf{I}_{F.} \\ \mathsf{I}_{FS} \\ \mathsf{I}_{.S} \end{bmatrix} \begin{bmatrix} \boxed{\mathbb{K}} \end{bmatrix} \begin{bmatrix} \quad \end{bmatrix} \begin{bmatrix} \mathsf{O}_S \\ \mathsf{O}_{FS} \\ \boxed{\mathbb{L}} \ \mathsf{O}_{F.} \end{bmatrix}$$
$$\tag{232}$$
$$\tag{233}$$

10 Markov variable maps and variables form a Markov category

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

- For every $A \in \mathcal{Y}$, the function $x \mapsto \mathbb{K}(x, A)$ is $\mathcal{X}$-measurable

- For every $x \in X$, the function $A \mapsto \mathbb{K}(x, A)$ 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

$$\delta(x, A) = \begin{cases} 1 & x \in A \\ 0 & \text{otherwise} \end{cases} \tag{234}$$

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

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

$$\text{del}_X := \quad \longrightarrow\!\!* \qquad \text{copy}_X \qquad := \quad \longrightarrow\!\!\!< \tag{235}$$

and satisfying the commutative comonoid equations

$$\tag{236}$$

$$\tag{237}$$

$$\tag{238}$$

as well as compatibility with the monoidal structure

$$\tag{239}$$

$$\tag{240}$$

and the naturality of *del*, which means that

$$\tag{241}$$

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

$$\mathbb{K}\mathbb{L} : \mathsf{X} \to \Delta(\mathsf{Z}) \tag{242}$$

$$\text{defined by } (\mathbb{K}\mathbb{L})(x, A) = \int_Y \mathbb{L}(y, A)\mathbb{K}(x, dy) \tag{243}$$

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

$$\mathbb{K} \otimes \mathbb{L} : (\mathsf{X}, \mathsf{W}) \to \Delta(\mathsf{Y}, \mathsf{Z}) \tag{244}$$

$$\text{defined by } \mathbb{K} \otimes \mathbb{L}(x, w, A \times B) = \mathbb{K}(x, A)\mathbb{L}(w, B) \tag{245}$$

52

The identity map is

$$\mathrm{Id}_{\mathsf{X}} : \mathsf{X} \to \Delta(\mathsf{X}) \tag{246}$$

$$\text{defined by}(\mathrm{Id}_X)(x, A) = \delta(x, A) \tag{247}$$

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

The swap map is given by

$$\mathrm{swap}_{\mathsf{X},\mathsf{Y}} : (\mathsf{X}, \mathsf{Y}) \to \Delta(\mathsf{Y}, \mathsf{X}) \tag{248}$$

$$\text{defined by}(\mathrm{swap}_{\mathsf{X},\mathsf{Y}})(x, y, A \times B) = \delta(x, B)\delta(y, A) \tag{249}$$

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

The copy map is given by

$$\mathrm{copy}_{\mathsf{X}} : \mathsf{X} \to \Delta(\mathsf{X}, \mathsf{X}) \tag{250}$$

$$\text{defined by}(\mathrm{copy}_X)(x, A \times B) = \delta_x(A)\delta_x(B) \tag{251}$$

and the erase map by

$$\mathrm{del}_{\mathsf{X}} : \mathsf{X} \to \Delta(*) \tag{252}$$

$$\text{defined by}(\mathrm{del}_X)(x, A) = \delta(*, A) \tag{253}$$

$$\tag{254}$$

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

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

Proof.

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

Equations 236 to 241 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.

$\square$

References

Ethan D. Bolker. Functions Resembling Quotients of Measures. *Transactions of the American Mathematical Society*, 124(2):292–312, 1966. ISSN 0002-9947. doi: 10.2307/1994401. URL https://www.jstor.org/stable/1994401. Publisher: American Mathematical Society.

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

Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, August 2019. ISSN 0960-1295, 1469-8072. doi: 10.1017/S0960129518000488.

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

A. Philip Dawid. Decision-theoretic foundations for statistical causality. *arXiv:2004.12493 [math, stat]*, April 2020. URL http://arxiv.org/abs/2004.12493. arXiv: 2004.12493.

M. P. Ershov. Extension of Measures and Stochastic Equations. *Theory of Probability & Its Applications*, 19(3):431–444, June 1975. ISSN 0040-585X. doi: 10.1137/1119053. URL https://epubs.siam.org/doi/abs/10.1137/1119053. Publisher: Society for Industrial and Applied Mathematics.

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

Brendan Fong. Causal Theories: A Categorical Perspective on Bayesian Networks. *arXiv:1301.6201 [math]*, January 2013. URL http://arxiv.org/abs/1301.6201. arXiv: 1301.6201.

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

D. Heckerman and R. Shachter. Decision-Theoretic Foundations for Causal Reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, December 1995. ISSN 1076-9757. doi: 10.1613/jair.202. URL https://www.jair.org/index.php/jair/article/view/10151.

M. A. Hernán and S. L. Taubman. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal*

of Obesity, 32(S3):S8–S14, August 2008. ISSN 1476-5497. doi: 10.1038/ijo. 2008.82. URL https://www.nature.com/articles/ijo200882.

Alan Hájek. What Conditional Probability Could Not Be. *Synthese*, 137(3): 273–323, December 2003. ISSN 1573-0964. doi: 10.1023/B:SYNT.0000004904. 91112.16. URL https://doi.org/10.1023/B:SYNT.0000004904.91112.16.

Bart Jacobs, Aleks Kissinger, and Fabio Zanasi. Causal Inference by String Diagram Surgery. In Mikoaj Bojaczyk and Alex Simpson, editors, *Foundations of Software Science and Computation Structures*, Lecture Notes in Computer Science, pages 313–329. Springer International Publishing, 2019. ISBN 978-3-030-17127-8.

Richard C. Jeffrey. *The Logic of Decision*. University of Chicago Press, July 1965. ISBN 978-0-226-39582-1.

James M. Joyce. *The Foundations of Causal Decision Theory*. Cambridge University Press, Cambridge ; New York, April 1999. ISBN 978-0-521-64164-7.

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

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

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

David K Lewis. Causation. *Journal of Philosophy*, 1986.

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.

Robert Nozick. Newcombs Problem and Two Principles of Choice. In Nicholas Rescher, editor, *Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of his Sixty-Fifth Birthday*, Synthese Library, pages 114–146. Springer Netherlands, Dordrecht, 1969. ISBN 978-94-017-1466-2. doi: 10.1007/978-94-017-1466-2_7. URL https://doi.org/10.1007/978-94-017-1466-2_7.

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.

Frank P. Ramsey. Truth and Probability. In Horacio Arló-Costa, Vincent F. Hendricks, and Johan van Benthem, editors, *Readings in Formal Epistemology: Sourcebook*, Springer Graduate Texts in Philosophy, pages 21–45. Springer International Publishing, Cham, 2016. ISBN 978-3-319-20451-2. doi: 10.1007/978-3-319-20451-2_3. URL https://doi.org/10.1007/978-3-319-20451-2_3.

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.

Leonard J. Savage. *The Foundations of Statistics*. Wiley Publications in Statistics, 1954.

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.

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

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

Paul Weirich. Causal Decision Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016. URL https://plato.stanford.edu/archives/win2016/entries/decision-causal/.

Appendix: