

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

David Johnston

December 3, 2021

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 2 |
| 2 | Variables and Probability Models | 3 |
| 2.1 | Section outline | 3 |
| 2.1.1 | Brief outline of probability gap models | 4 |
| 2.2 | Probability distributions, Markov kernels and string diagrams . . | 5 |
| 2.2.1 | Examples | 7 |
| 2.2.2 | Example: comb insertion | 8 |
| 2.3 | Semantics of observed and unobserved variables | 9 |
| 2.4 | Events | 12 |
| 2.5 | Probabilistic models for causal inference | 13 |
| 2.6 | Probability gaps | 14 |
| 2.7 | Probability gap models defined by marginal and conditional probabilities | 17 |
| 2.8 | Higher order gap models | 19 |
| 2.9 | Order 2 gaps: probability combs | 21 |
| 2.10 | Revisiting truncated factorisation | 24 |
| 2.10.1 | Disintegrations | 25 |
| 2.10.2 | Conditional independence | 26 |
| 2.11 | Recursive disintegration | 29 |
| 2.11.1 | Graphical properties of conditional independence | 29 |
| 2.11.2 | Restricted 2-combs | 30 |
| 2.12 | Results I use that don't really fit into the flow of the text | 31 |
| 2.12.1 | Repeated variables | 31 |
| 3 | Decision theoretic causal inference | 33 |
| 3.1 | Decision problems | 34 |
| 3.2 | What should a probability model represent? Controversies about decision theories | 35 |
| 3.3 | Unresponsiveness | 36 |
| 3.4 | Causal models similar to see-do models | 36 |

| | | |
|-----------|--|-----------|
| 3.5 | See-do models and classical statistics | 37 |
| 4 | Causal Bayesian Networks | 38 |
| 4.1 | Probability 2-combs represented by causal Bayesian networks . . | 39 |
| 4.2 | See-do models compatible with causal Bayesian networks | 42 |
| 4.3 | Proxy control | 44 |
| 5 | Potential outcomes | 45 |
| 6 | Appendix: see-do model representation | 49 |
| 7 | Appendix: Counterfactual representation | 51 |
| 7.1 | Parallel potential outcomes representation theorem | 52 |
| 8 | Appendix: Connection is associative | 55 |
| 9 | Appendix: String Diagram Examples | 56 |
| 10 | Markov variable maps and variables form a Markov category | 57 |

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 index¹ are vaguely

¹the 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.

defined, potential outcome variables and causal effects themselves become ill-defined. We note that “actions targeting body mass index” are not elements of a potential outcomes model but “things to which potential outcomes should correspond”. The authors claim is that vagueness in the “territory” leads to ambiguity about elements of the “map” – and, as we have suggested, anything we can try to say about the territory is unavoidably vague. This seems like a serious problem.

In a response, Pearl (2018) argues that *interventions* (by which we mean the operation defined by a causal graphical model) are well defined, but may not always be a good model of an action. Pearl further suggests that interventions in graphical models correspond to “virtual interventions” or “ideal, atomic interventions”, and that perhaps carefully chosen interventions can be good models of actions. Shahar (2009), also in response, argued that interventions targeting body mass index applied to correctly specified graphical causal models will necessarily yield no effect on anything else which, together with Pearl’s suggestion, implies perhaps that an “ideal, atomic intervention” on body mass index cannot have any effect on anything else. If this is so, it seems that we are dealing with quite a serious case of vagueness – there is a whole body of literature devoted to estimating a “causal effect” that, it is claimed, is necessarily equal to zero! Authors of the original literature on the effects of BMI might counter that they were estimating something different that wasn’t necessarily zero, but as far as we are concerned such a response would only underscore the problem of ambiguity.

One of the key problems in this whole discussion is how the things we have 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 Section outline

This section introduces the mathematical foundations used throughout the rest of the paper. The first subsection briefly introduces probability theory, which is likely to be familiar to many readers, as well as how string diagrams can be used to represent probabilistic functions (or *Markov kernels*), which may be less familiar. We use string diagrams for probabilistic reasoning in a number of places, and this section is intended to help interpret mathematical statements in this form.

The second subsection discusses the interpretation of probabilistic variables. Our formalisation of probabilistic variables is standard – we define them as measurable functions on a fundamental probability set Ω . We discuss how this formalisation can be connected to statements about the real world via *measurement processes*, and distinguishes observed variables (which are associated with measurement processes) from unobserved variables (which are not associated with measurement processes). This section is not part of the mathematical theory of probability gap models, but it is relevant when one wants to apply this theory to real problems or to understand how the theory of probability gap models relates to other theories of causal inference.

Finally, we introduce *probability gap models*. Probability gap models are a generalisation of probability models, and to understand the rest of this paper a reader needs to understand what a probability gap model is, how we define the common kinds of probability gap models used in this paper and what conditional probabilities and conditional independence statements mean for probability gap models.

2.1.1 Brief outline of probability gap models

We consider a probability model to be a probability space $(\Omega, \mathcal{F}, \mu)$ along with a collection of random variables. However, if I want to use probabilistic models to support decision making, then I need function from options to probability models. For example, suppose I have two options $A = \{0, 1\}$, and I want to compare these options based on what I expect to happen if I choose them. If I choose option 0, then I can (perhaps) represent my expectations about the consequences with a probability model, and if I choose option 1 I can represent my expectations about the consequences with a different probability model. I can compare the two consequences, then decide which option seems to be better. To make this comparison, I have used a function from elements of A to probability models. A function that takes elements of some set as inputs (which may or may not be decisions) and returns probability models is a *probability gap model*, and the set of inputs it accepts is a *probability gap*.

We are particularly interested in probability gap models where the consequences of all inputs share some marginal or conditional probabilities. The simplest example of a model like this can be represented by a probability distribution \mathbb{P}^X for some variable $X : \Omega \rightarrow X$. Such a probability distribution is consistent with many base measures on the fundamental probability set Ω , and so we can consider the choice of base measure to be a probability gap. Not every probability distribution over X can define a probability gap model in this way. In particular, we need \mathbb{P}^X to assign probability 0 to outcomes that are mathematically impossible according to the definition of X to ensure that there is some base measure that features \mathbb{P}^X as a marginal. We call probability gap models represented by probability distributions *order 0 probability gap models*.

Higher order probability gap models can be represented by conditional probabilities $\mathbb{P}^{Y|X}$ or pairs of conditional probabilities $\{\mathbb{P}^{X|W}, \mathbb{P}^{Z|WXY}\}$, which we call *order 1* and *order 2* models respectively. Decision functions in data-

driven decision problems correspond to probability gaps in order 2 models, as we discuss in Section 3, which makes this type of model particularly interesting for our purposes. We also require these to be valid, and we define conditions for validity and prove that they are sufficient to ensure that models represented by conditional probabilities can in fact be mapped to base measures on the fundamental probability set.

A conditional independence statement in a probability gap model means that the corresponding conditional independence statement holds for all base measures in the range of the function defined by the model. It is possible to deduce conditional independences from “independences” in the conditional probabilities that we use to represent these models, and conditional independences can imply the existence of conditional probabilities with certain independence properties.

We can consider causal Bayesian networks to represent order 2 probability gap models. That is, a causal Bayesian network represents a function \mathbb{P} that takes inserts from some set A of conditional probabilities and returns a probability model, and it does so in such a way that there are a pair of conditional probabilities $\{\mathbb{P}^{X|W}, \mathbb{P}^{Z|WXY}\}$ shared by all models in the codomain of \mathbb{P} . The observational distribution is the value of $\mathbb{P}(\text{obs})$ for some *observational insert* $\text{obs} \in A$, and other choices of inserts yield interventional distributions. Defining causal Bayesian networks in this manner resolves two areas of difficulty with causal Bayesian networks. First, under the standard definition of causal Bayesian networks interventional probabilities may fail to exist; with our perspective we can see that this arises due to misunderstanding the domain of \mathbb{P} . Secondly, there may be multiple distributions that differ in important ways that all satisfy the standard definition of “interventional distributions”. The one-to-many relationship between observations and interventions is a basic challenge of causal inference, the problem arises when this relationship is obscured by calling multiple different things “the interventional distribution”. If we consider causal Bayesian networks to represent order 2 probability gap models, we avoid doing this.

2.2 Probability distributions, Markov kernels and string diagrams

We make use of a string diagram notation for probabilistic reasoning. Graphical models are often employed in causal reasoning, and string diagrams are a particularly rigorous graphical notation for probabilistic models. It 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 valid 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 applicable.

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

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

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:

$$\mathbb{K} := \boxed{\mathbb{K}} \quad (1)$$

$$\mathbb{P} := \triangleleft \mathbb{P} \quad (2)$$

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

$$\mathbb{LM} := X \boxed{\mathbb{K}} \boxed{\mathbb{M}} Z \quad (3)$$

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 $(\mathbb{K} \otimes \mathbb{L})(y, z | w, x) := \mathbb{K}(y|w)\mathbb{L}(z|x)$. The tensor product is graphically represented by drawing kernels in parallel:

$$\mathbb{K} \otimes \mathbb{L} := \begin{array}{c} W \boxed{\mathbb{K}} Y \\ X \boxed{\mathbb{L}} Z \end{array} \quad (4)$$

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 $\text{id}_X : X \rightarrow X$ defined by $(\text{id}_X)(x'|x) = \llbracket x = x' \rrbracket$, where the Iverson bracket $\llbracket \cdot \rrbracket$ evaluates to 1 if \cdot is true and 0 otherwise, is a bare line:

$$\text{id}_X := X \text{---} X \quad (5)$$

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 $\text{del}_X : X \rightarrow \{*\}$ defined by $(\text{del}_X)(*|x) = 1$ is a Markov kernel that “discards the input”. It is drawn as a fuse:

$$\text{del}_X := \text{---} * \text{---} X \quad (6)$$

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

$$\text{copy}_X := X \text{---} \begin{array}{c} \diagup \\ \diagdown \end{array} \begin{array}{c} X \\ X \end{array} \quad (7)$$

The swap map $\text{swap}_{X,Y} : X \times Y \rightarrow Y \times X$ defined by $(\text{swap}_{X,Y})(y', x'|x, y) = \llbracket x = x' \rrbracket \llbracket y = y' \rrbracket$ swaps two inputs, and is represented by crossing wires:

$$\text{swap}_X := \begin{array}{c} \diagup \quad \diagdown \\ \diagdown \quad \diagup \end{array} \quad (8)$$

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

2.2.1 Examples

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

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

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

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

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

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

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

$$= \mathbb{K}(A|x, y, y) \quad (12)$$

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

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

$$\text{copy}_Y^n := \begin{cases} \text{copy}_Y^{n-1}(\text{id}_{Y^{n-2}} \otimes \text{copy}_Y) & n > 2 \\ \text{copy}_Y & n = 2 \end{cases} \quad (13)$$

$$(\text{copy}_Y^2 \mathbb{K}^2)(z|y) = \mathbb{K}^2(z|y, y) \quad (14)$$

$$(15)$$

Suppose for induction

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

then

$$(\text{copy}_Y^n \mathbb{K}^n)(z|y) = (\text{copy}_Y^{n-1}(\text{id}_{Y^{n-2}} \otimes \text{copy}_Y) \mathbb{K}^n)(z|y) \quad (17)$$

$$= \sum_{y' \in Y^{n-1}} (\text{id}_{Y^{n-2}} \otimes \text{copy}_Y)(\mathbf{y}'|(y)^{n-1}) \mathbb{K}^n(z|\mathbf{y}') \quad (18)$$

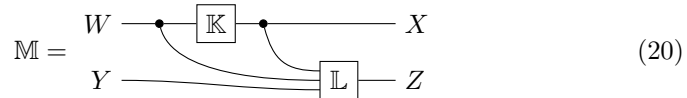
$$= \mathbb{K}^n(z|(y)^n) \quad (19)$$

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.

2.2.2 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 $\mathbb{M} : W \times Y \rightarrow X \times Z$ defined as



$$\mathbb{M} = \quad (20)$$

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

$$\mathbb{M}(A \times B|w, y) = [(\text{copy}_W \otimes \text{id}_Y)(\mathbb{K} \otimes \text{id}_{W \times Y}) \quad (21)$$

$$(\text{copy}_X \otimes \text{id}_{W \times Y})(\text{id}_X \otimes \mathbb{L})](A \times B|w, y) \quad (22)$$

$$= [(\mathbb{K} \otimes \text{id}_{W \times Y})(\text{copy}_X \otimes \text{id}_{W \times Y}) \quad (23)$$

$$(\text{id}_X \otimes \mathbb{L})](A \times B|w, w, y) \quad (24)$$

$$= \int_X (\text{id}_X \otimes \mathbb{L})(A \times B|x', w, y) \mathbb{K}(dx'|w)(y, z|y', x) \quad (25)$$

$$= \int_X \text{id}_X(A|x') \mathbb{L}(B|x', w, y) \mathbb{K}(dx'|w) \quad (26)$$

$$= \int_A \mathbb{L}(B|x', w, y) \mathbb{K}(dx'|w) \quad (27)$$

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} :

$$\mathbb{N} = W \text{ --- } \bullet \text{ --- } \boxed{\mathbb{K}} \text{ --- } \bullet \text{ --- } \boxed{\mathbb{J}} \text{ --- } \bullet \text{ --- } \boxed{\mathbb{L}} \text{ --- } \begin{matrix} Z \\ Y \\ X \end{matrix} \quad (28)$$

We can translate Equation 28 to

$$\mathbb{N}(A \times B \times C|w) = [\text{copy}_W(\mathbb{K} \text{copy}_Y^3 \otimes \text{id}_W) \quad (29)$$

$$(\text{id}_Y \otimes \mathbb{J} \otimes \text{id}_Y)(\text{id}_Y \otimes \text{copy}_X \otimes \text{id}_Y) \quad (30)$$

$$(\mathbb{L} \otimes \text{id}_X \otimes \text{id}_Y)](A \times B \times C|w) \quad (31)$$

$$= [(\mathbb{K} \text{copy}_Y^3 \otimes \text{id}_W)(\text{id}_Y \otimes \mathbb{J} \otimes \text{id}_Y) \quad (32)$$

$$(\text{id}_Y \otimes \text{copy}_X \otimes \text{id}_Y) \quad (33)$$

$$(\mathbb{L} \otimes \text{id}_X \otimes \text{id}_Y)](A \times B \times C|w, w) \quad (34)$$

$$= \int_X \int_Y \mathbb{L}(C|x', w, y') \text{id}_X(A|x') \text{id}_Y(B|y') \mathbb{J}(dy'|x', w) \mathbb{K}(dx'|w) \quad (35)$$

$$= \int_A \int_B \mathbb{L}(C|x', w, y') \mathbb{J}(dy'|x', w) \mathbb{K}(dx'|w) \quad (36)$$

2.3 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 and 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 fundamental probability set 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 peculiar. 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 $\bigvee_{x \in A} \mathcal{B} \bowtie x$. 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 $f : B \rightarrow C$, 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 $x \in B$:

$$(g \circ f) \circ \mathcal{B} \text{ yields } x \iff B \text{ yields } (g \circ f)^{-1}(x) \quad (37)$$

$$\iff B \text{ yields } f^{-1}(g^{-1}(x)) \quad (38)$$

$$\iff f \circ B \text{ yields } g^{-1}(x) \quad (39)$$

$$\iff g \circ (f \circ B) \text{ yields } x \quad (40)$$

One might wonder 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 $\mathcal{F} = \mathcal{M}\mathcal{A}$, 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 $\mathcal{M}\mathcal{A} = h \circ (\mathcal{M}, \mathcal{A}) = (h \circ g) \circ \mathcal{S}$. Note that each procedure is associated with a unique function with domain Ψ .

Thus far, Ψ is a “fundamental probability set” that only contains observable variables. To include unobserved variables, we posit a richer fundamental probability set 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 fundamental probability set Ω such that $|\Omega| \geq |\Psi|$ and a surjective function $\mathbf{S} : \Omega \rightarrow \Psi$ associated with \mathcal{S} . We connect Ω , \mathbf{S} and \mathcal{S} with the notion of *consistency with observation*:

$$\omega \in \Omega \text{ is consistent with observation iff the result yielded by } \mathcal{S} \text{ is equal to } \mathbf{S}(\omega) \quad (41)$$

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 $\mathbf{S}^{-1}(s)$.

One thing to note in this setup is that two different sets of measurement outcomes Ψ and Ψ' entail a different measurement procedures \mathcal{S} and \mathcal{S}' , but different fundamental probability sets 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 fundamental probability set (e.g., the set of elementary events) to the real line. In our definition, the mapping is from the fundamental probability set 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 $f \circ S$ for some f – is modeled by a variable, but there may be variables defined on Ω that do not factorise through S , and these variables do not model procedures.

2.4 Events

To recap, we have a procedure S yielding values in Ψ that measures everything we are interested in, a fundamental probability set Ω and a function S that models 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 $\sigma(S) \subset \mathcal{F}$. If a procedure $\mathcal{X} = f \circ S$ then we define $X : \Omega \rightarrow X$ by $X := f \circ S$.

If a particular procedure $\mathcal{X} = f \circ S$ eventually yields a value x , then the values of Ω consistent with observation must be a subset of $X^{-1}(x)$. We define an *event* $X \bowtie x := X^{-1}(x)$, which we read “the event that X yields x ”. An event $X \bowtie x$ occurs if the consistent values of Ω are a subset of $X \bowtie x$, 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 $A \in \mathcal{X}$, $X \bowtie A = \bigcup_{x \in A} X \bowtie x$.

Given $Y : \Omega \rightarrow X$, we can define a sequence of variables: $(X, Y) := \omega \mapsto (X(\omega), Y(\omega))$. (X, Y) has the property that $(X, Y) \bowtie (x, y) = X \bowtie x \cap Y \bowtie y$, 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 $Y = y$ already has a meaning, namely that Y is a constant function everywhere equal to y .

2.5 Probabilistic models for causal inference

The fundamental probability set (Ω, \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”.

Given a model skeleton, a common approach to attaching a probabilistic model involves defining a base measure μ on Ω which yields a probability space $(\Omega, \mathcal{F}, \mu)$. For causal inference, we need a to generalise this approach, because we need to handle *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 or “inserts”. If we complete such a model with an appropriate insert, we get a standard probability model.

Probability gap models are particularly useful in for decision making. 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 I need a model that can take a provisional choice as an argument and return a probability model representing the results of this choice; in other words, the choices I may make are *probability gaps*.

We will consider a motivating example initially posed using the language of causal Bayesian networks. For this example, we will assume that the reader is familiar enough with causal Bayesian networks to follow along. We will offer more careful definitions later.

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

$$\mathbb{P}_x^{XYZ}(x', y, z) = \bar{\mathbb{P}}^{Y|XZ}(y|x, z) \mathbb{P}^Z(z) \llbracket x = x' \rrbracket \quad (42)$$

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; $\mathbb{P}^Z = \mathbb{P}_x^Z$. Finally, there must be some version of the interventional conditional probability $Y|(X, Z)$ that is equal to some version of the observational conditional probability $Y|(X, Z)$; there exists $\bar{\mathbb{P}}^{Y|XZ}$ and $\bar{\mathbb{P}}_x^{Y|XZ}$ such that $\bar{\mathbb{P}}^{Y|XZ} = \bar{\mathbb{P}}_x^{Y|XZ}$. We use the overbars here to indicate that, unlike in familiar cases, the particular choice of $\bar{\mathbb{P}}^{Y|XZ}$ can matter here.

The operation of “setting X to x ” is often referred to as an *intervention*. Interventions are things we can choose to do, or not to do. We can also consider choosing to do or not do an intervention based on the output of some random process. We need some kind of model that can tell us which result we are likely to see for any choice of interventions, random or nonrandom. This means that we need a model with a *probability gap* for the choice of interventions. For a nonrandom choice of intervention x , we can consider the map $x \mapsto \mathbb{P}_x^{XYZ}$ such a model, and if we include random choices we can consider $\mathbb{Q} : \mathbb{P}_\alpha^X \mapsto \sum_x \mathbb{P}_\alpha^X(x) \mathbb{P}_x^{XYZ}$ to be such a model.

\mathbb{Q} , as we have defined it, is not quite an ideal candidate for a probability gap model. Firstly, the conditional probability $\bar{\mathbb{P}}^{Y|XZ}$ may be chosen arbitrarily on a set of measure zero with regard to \mathbb{P}^{XZ} . As a result, depending on the choice of $\bar{\mathbb{P}}^{Y|XZ}$, Equation 42 can be satisfied by multiple probability distributions that differ in meaningful ways. For example, suppose X , Y and Z are binary and $\mathbb{P}((X, Z) \bowtie (1, 1)) = 1$. Then we can consistently choose $\mathbb{P}^{Y|XZ}(1|0, 1) = 1$ or $\mathbb{P}^{Y|XZ}(1|0, 1) = 0$ because $\{0, 1\}$ is a measure zero event. However, the first choice gives us $\mathbb{P}_0^{XYZ}(0, 1, 1) = 1$ while the second gives us $\mathbb{P}_0^{XYZ}(0, 1, 1) = 0$, 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 $X = f \circ Z$ for some f . Then we must have $\mathbb{P}_x^X(x') = \mathbb{P}_x^Z(f^{-1}(x'))$ for any x . However, we also have $\mathbb{P}_x^X(x') = \llbracket x = x' \rrbracket$ for all x, x' and $\mathbb{P}_x^Z = \mathbb{P}^Z$ 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.

A more subtle example of this latter problem appears in Shahar (2009). A causal graph in that paper features an arrow $Z \longrightarrow X$ where $Z = (H, W)$, representing a person’s height and weight, and X represents their body mass index. This causal model is used to draw conclusions about the result of intervening on X . By definition, $X = \frac{W}{H^2}$. While we don’t have X equal to Z , it must still be a deterministic function of Z . However, any intervention on X along the lines of Equation 42 will yield X independent of (H, W) , and unless (H, W) is deterministically equal to a constant and the intervention on X is carefully chosen, there is no probability model at all that has this independence.

The theory of probability gap models allows us to model things like interventions and it does not share these problems of non-uniqueness and non-existence with models defined via truncated factorisation.

2.6 Probability gaps

A probability gap model is a function that maps “inserts” to probability models. We think of the set of inserts as different ways we can fill the probability gap. The particular kinds of inserts we want to consider here are marginal probabilities and conditional probabilities.

Definition 2.3 (Probability gap model). Given a fundamental probability set Ω and a set of inserts A , a probability gap model $\mathbb{P} : A \rightarrow \Delta(\Omega)$ is a function

that sends an element of A to a probability measure on Ω , which we call a *base measure*.

We will make a substantial simplifying assumption: all sets, including the fundamental probability set 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: $\mathbb{P}^X(x) := \mathbb{P}^X(\{x\})$.

Definition 2.4 (Probability space). A probability space is a triple $(\mu, \Omega, \mathcal{F})$, where μ is a base measure on \mathcal{F} which we here take to be $\mathcal{P}(\Omega)$.

Probability spaces along with random variables can be used to define *marginal probability distributions* of those random variables.

Definition 2.5 (Marginal distribution with respect to a probability space). Given a probability space (μ, Ω) and a random variable $X : \Omega \rightarrow X$, we can define the *marginal distribution* of X with respect to μ , $\mu^X : \mathcal{X} \rightarrow [0, 1]$ of X by $\mu^X(x) := \mu(X \bowtie x)$ for any $x \in X$. Equivalently, if we define the Markov kernel $\mathbb{F}_X : \Omega \rightarrow X$ associated with the function X by $\mathbb{F}_X(x|\omega) = \llbracket X(\omega) = x \rrbracket$, then

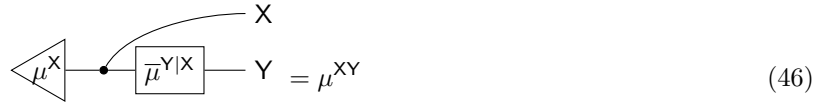
$$\mu^X = \mu \mathbb{F}_X \quad (43)$$

A $Y|X$ -conditional probability is “the probability of Y given X relative to μ ”. It is a Markov kernel that maps the marginal distribution of X to the marginal distribution of the sequence (X, Y) .

Definition 2.6 (Conditional probability with respect to a probability space). Given a probability space (μ, Ω) and random variables $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, the disintegration $\mu^{Y|X} : X \rightarrow Y$ is any Markov kernel such that

$$\mu^X(x) \mu^{Y|X}(y|x) = \mu^{XY}(x, y) \quad \forall x \in X, y \in Y \quad (44)$$

$$\iff \quad (45)$$



$$\quad (46)$$

Because in general disintegrations are non-unique and we sometimes need to account for this fact, we use a bar over the top of the letter to indicate that it is an arbitrary element of the set of disintegrations.

Given a probability gap model \mathbb{P} , we define marginal and conditional probabilities as probability measures and Markov kernels that satisfy Definitions 2.5 and 2.6 respectively for *all* base measures in the range of \mathbb{P} .

Even though there are generally multiple Markov kernels that satisfy the properties of a conditional probability with respect to a probability gap model, this definition ensures that any choice will map the same marginal distribution to the same joint distribution *no matter which insert we choose*. This is different

to the example of truncated factorisation we started with, where different choices of $\bar{\mathbb{P}}^{Y|XZ}$ gave different joint probabilities for some insert choices. This is because $\bar{\mathbb{P}}^{Y|XZ}$ is a conditional probability with respect to the distribution of observations, and it is not a conditional probability with respect to the probability gap model of interventions.

Definition 2.7 (Marginal probability with respect to a probability gap model). Given a fundamental probability set Ω , a set of inserts A , a variable $X : \Omega \rightarrow X$ and a probability gap model $\mathbb{P} : A \rightarrow \Delta(\Omega)$, if $\mathbb{P}_a^X = \mathbb{P}_{a'}^X$ for all $a, a' \in A$, the marginal distribution $\mathbb{P}^X = \mathbb{P}_a^X$ for any a . Otherwise, it is undefined.

Definition 2.8 (Conditional probability with respect to a probability gap model). Given a fundamental probability set Ω , a set of inserts A , variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$ and a probability gap model $\mathbb{P} : A \rightarrow \Delta(\Omega)$, $\mathbb{P}^{Y|X}$ is any Markov kernel $X \rightarrow Y$ such that for all $\mathbb{P}_a := \mathbb{P}(a)$, $a \in A$

$$\mathbb{P}_a^X(x) \mathbb{P}^{Y|X}(y|x) = \mathbb{P}_a^{XY}(x, y) \quad \forall x \in X, y \in Y \quad (47)$$

If no such Markov kernel exists, $\mathbb{P}^{Y|X}$ is undefined.

Given a conditional probability with respect to a probability gap model, we can find other conditional probabilities by “pushing forward” via a function.

Theorem 2.9 (Recursive pushforward). *Suppose we have a fundamental probability set Ω , a set of inserts A , variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$, $Z : \Omega \rightarrow Z$ and a probability gap model $\mathbb{P} : A \rightarrow \Delta(\Omega)$ such that $\mathbb{P}^{X|Y}$ is a $Y|X$ conditional probability of \mathbb{P} and $Z = f \circ Y$ for some $f : Y \rightarrow Z$. Then there exists a $Z|X$ conditional probability of \mathbb{P} defined by $\mathbb{P}^{Z|X}(z|x) := \mathbb{P}^{Y|X}(f^{-1}(z)|x)$.*

Proof. For any $a \in A$, x, z

$$\mathbb{P}_a^X(x) \mathbb{P}_a^{Z|X}(z|x) = \mathbb{P}_a(X^{-1}(x) \cap Z^{-1}(z)) \quad (48)$$

$$= \mathbb{P}_a(X^{-1}(x) \cap Y^{-1}(f^{-1}(z))) \quad (49)$$

$$= \mathbb{P}_a^{X,Y}(\{x\} \times f^{-1}(z)) \quad (50)$$

$$= \mathbb{P}_a^X(x) \mathbb{P}_a^{Y|X}(f^{-1}(z)|x) \quad (51)$$

□

Given a conditional probability with respect to a probability gap model, we can also find additional conditional probabilities by disintegrating the original conditional probability.

Copy-paste the relevant theorems

Theorem 2.10 (Recursive disintegration). *Suppose we have a fundamental probability set Ω , a set of inserts A , variables $W : \Omega \rightarrow W$, $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$, $Z : \Omega \rightarrow Z$ and a probability gap model $\mathbb{P} : A \rightarrow \Delta(\Omega)$ such that $\mathbb{P}^{X|Y}$ is a $Y|X$ disintegration of \mathbb{P} . Define an order 1 probability gap model $\mathbb{Q}^{\square Y|X}$ such that $\mathbb{Q}^{Y|X} = \mathbb{P}^{Y|X}$. Then if $\mathbb{Q}^{Z|W}$ is a $Z|W$ disintegration of \mathbb{Q} , it is also a $Z|W$ disintegration of \mathbb{P} .*

Proof. The range of \mathbb{Q} is all base measures on Ω for which $\mathbb{P}^{\mathbf{X}|\mathbf{Y}}$ is a $\mathbf{Y}|\mathbf{X}$ disintegration. By assumption, then $\text{Range}(\mathbb{Q}) \supset \text{Range}(\mathbb{P})$. By definition, $\mathbb{Q}^{\mathbf{Z}|\mathbf{W}}$ is a $\mathbf{Z}|\mathbf{W}$ disintegration of every base measure in $\text{Range}(\mathbb{Q})$ and so it is also a $\mathbf{Z}|\mathbf{W}$ disintegration of every base measure in $\text{Range}(\mathbb{P})$. \square

2.7 Probability gap models defined by marginal and conditional probabilities

In the previous section we defined marginal and conditional probabilities for probability gap models. We can in turn define probability gap models by specifying a collection of marginal and/or conditional probabilities along with a set of inserts. In fact, we will define a hierarchy of probability gap models – order 0 gap models are specified by marginal probabilities and map inserts to base measures, while order 1 gap models are specified by conditional probabilities and map inserts to order 0 gap models.

It is helpful to define an operation \odot that combines Markov kernels in the same that Definition 2.6 combined marginals and conditionals to a joint probability.

Definition 2.11 (Copy-product). Given two Markov kernels $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : Y \times X \rightarrow Z$, define the copy-product $\mathbb{K} \odot \mathbb{L} : X \rightarrow Y \times Z$ as

$$\mathbb{K} \odot \mathbb{L} := \text{copy}_X(\mathbb{K} \otimes \text{id}_X)(\text{copy}_Y \otimes \text{id}_X)(\text{id}_Y \otimes \mathbb{L}) \quad (52)$$

$$= \begin{array}{c} \text{Diagram showing the composition of Markov kernels } \mathbb{K} \text{ and } \mathbb{L} \text{ to form } \mathbb{K} \odot \mathbb{L}. \\ \text{A horizontal line from } X \text{ splits into two paths. The top path goes through a box labeled } \mathbb{K} \text{ to } Y. \\ \text{The bottom path goes through a box labeled } \mathbb{L} \text{ to } Z. \\ \text{A curved arrow connects the output of } \mathbb{K} \text{ back to the input of } \mathbb{L}, \text{ representing the copy operation.} \end{array} \quad (53)$$

$$\iff \quad (54)$$

$$(\mathbb{K} \odot \mathbb{L})(y, z|x) = \mathbb{L}(z|y, x)\mathbb{K}(y|x) \quad (55)$$

We will first consider “order 0” probability gap models defined by marginal probabilities. These serve as a base kind of probability gap model that we use to define higher order models. The “probability gap” in this case relates to the particular choice of base measure that yields this marginal probability. Like most statistical work, we don’t typically explicitly define the fundamental probability set Ω , nor do we explicitly care about the base measure on it. Thus, unlike higher order probability gap models, we’re not particularly interested in the probability gaps in order 0 models. Nevertheless, they serve as a base type for the purposes of defining higher order models.

We define an order 0 probability gap model \mathbb{P} to be a function from some set of Markov kernels $X \rightarrow \Omega$ to probability measures on Ω such that for some $\mathbb{K} : X \rightarrow \Omega$, $\mathbb{P}(\mathbb{K}) = \mathbb{P}^{\mathbf{X}}\mathbb{K}$ for some probability distribution $\mathbb{P}^{\mathbf{X}}$. Not every Markov kernel $X \rightarrow \Omega$ will yield a base measure on Ω that pushes forward to $\mathbb{P}^{\mathbf{X}}$ by this operation. For example, let $X = \Omega = \{0, 1\}$ and $\mathbf{X} = \text{id}$, the identity function on Ω , $\mathbb{P}^{\mathbf{X}}(x) = \llbracket x = 1 \rrbracket$ and $\mathbb{K}(\omega|x) = \llbracket 1 - x \rrbracket$. Then letting $\mu := (\mathbb{P}^{\mathbf{X}}\mathbb{K})$,

$\mu^{\mathbf{X}}(x) = \mu(x) = \llbracket x = 0 \rrbracket$ which is not equal to $\mathbb{P}^{\mathbf{X}}(x)$. Instead we define \mathbb{P} on the set of *valid candidate conditionals*, which is a subset of Markov kernels $X \rightarrow \Omega$.

Not all probability measures $\mathbb{P}^{\mathbf{X}}$ on X map to base measures μ via a valid candidate conditional that yield back $\mathbb{P}^{\mathbf{X}}$ as marginals. Consider, for example, $\mathbf{X} = (\mathbf{Y}, \mathbf{Y})$ for some \mathbf{Y} and any measure $\mathbb{P}^{\mathbf{Y}\mathbf{Y}}$ that assigns nonzero probability to $(\mathbf{Y}, \mathbf{Y}) \bowtie (y, y')$ for $y \neq y'$. Then there is no base measure that pushes forward to $\mathbb{P}^{\mathbf{Y}\mathbf{Y}}$. Thus we restrict our attention to *valid candidate distributions* that can serve as definitions for order 0 probability gaps.

Definition 2.12 (Valid candidate distribution). A valid \mathbf{X} probability distribution $\mathbb{P}^{\mathbf{X}}$ is any probability measure on $\Delta(X)$ such that $\mathbf{X}^{-1}(x) = \emptyset \implies \mathbb{P}^{\mathbf{X}}(x) = 0$.

Definition 2.13 (Valid candidate conditional). Given (Ω, \mathcal{F}) , $\mathbf{X} : \Omega \rightarrow X$, $\mathbf{Y} : \Omega \rightarrow Y$ a *valid $\mathbf{Y}|\mathbf{X}$ conditional probability* $\mathbb{P}^{\mathbf{Y}|\mathbf{X}}$ is a Markov kernel $X \rightarrow Y$ such that it assigns probability 0 to contradictions:

$$\forall y \in Y, x \in X : (\mathbf{X}, \mathbf{Y}) \bowtie (x, y) = \emptyset \implies \left(\mathbb{P}^{\mathbf{Y}|\mathbf{X}}(y|x) = 0 \right) \vee (\mathbf{X} \bowtie (x) = \emptyset) \quad (56)$$

The definition of conditional probability (Definition 2.6) allows in some cases for conditional probabilities that are not valid candidate conditionals. In Theorem 2.30 we prove that it is always possible to choose a version of a conditional probability that is also a valid candidate conditional. See also Hájek (2003) for an argument that conditional probabilities should always satisfy the property of being valid candidate conditionals.

With valid candidate distributions and valid candidate conditionals, we can define order 0 probability gap models.

Definition 2.14 (Order 0 probability gap model). Given probability set Ω and variable $\mathbf{X} : \Omega \rightarrow X$, define the valid candidate conditionals $A \subset \Delta(\Omega)^X$. An order 0 probability gap model is a function $\mathbb{P} : A \rightarrow \Delta(\Omega)$ such that some $\mathbb{P}^{\mathbf{X}} \in \Delta(X)$ is a marginal distribution with respect to \mathbb{P} .

Definition 2.15 (Order 0 model associated with a given marginal). Given probability set Ω and variable $\mathbf{X} : \Omega \rightarrow X$, valid candidate conditionals $A \subset \Delta(\Omega)^X$ and a valid candidate distribution $\mathbb{P}^{\mathbf{X}} \in \Delta(X)$, define $\mathbb{P} : A \rightarrow \Delta(\Omega)$ by $\mathbb{K} \in A$, \mathbb{P} is given by

$$\mathbb{P} : \mathbb{K} \mapsto \mathbb{P}^{\mathbf{X}} \mathbb{K} \quad (57)$$

for any valid candidate conditional $\mathbb{K} \in A$.

Theorem 2.16 shows that Definition 2.15 does actually define an order 0 model with marginal $\mathbb{P}^{\mathbf{X}}$.

Theorem 2.16 (Completion). *Given $\mathbf{X} : \Omega \rightarrow X$, $\mathbb{J} \in \Delta(X)$ and $\mathbb{K} : X \rightarrow \Omega$, there exists some $\mu \in \Delta(\Omega)$ and $\mathbf{Y} : \Omega \rightarrow Y$ such that $\mathbb{J} \odot \mathbb{K} = \mu^{\mathbf{Y}}$ and $\mu^{\mathbf{X}} = \mathbb{J}$ if \mathbb{J} is a valid candidate distribution and \mathbb{K} is a valid candidate conditional, and only if \mathbb{J} is a valid candidate distribution.*

Proof. If: Define μ by

$$\mu(\omega) = \sum_{x \in X} (\mathbb{J} \odot \mathbb{K})(x, \omega) \quad (58)$$

Then

$$\mu^X(x) = \mu(X \bowtie x) \quad (59)$$

$$= \sum_{x' \in X} \mathbb{J}(x') \mathbb{K}(X \bowtie x | x') \quad (60)$$

$$= \sum_{x' \in X} \mathbb{J}(x') \mathbb{I}[x = x'] \quad \text{if } X \bowtie x \neq \emptyset \quad (61)$$

$$= \mathbb{J}(x) \quad (62)$$

and, letting $Y = (X, \mathbb{I})$, for any $x \in X$, $\omega \in \Omega$

$$\mu^Y(x, \omega) = \mu^{X\mathbb{I}}(x, \omega) \quad (63)$$

$$= \mu(\omega) \mu^{X\mathbb{I}}(x | \omega) \quad (64)$$

$$= \sum_{x' \in X} (\mathbb{J} \odot \mathbb{K})(x', \omega) \mathbb{I}[X(\omega) = x] \quad (65)$$

$$= (\mathbb{J} \odot \mathbb{K})(x, \omega) \quad (66)$$

$$\iff \quad (67)$$

$$\mathbb{J} \odot \mathbb{K} = \mu^Y \quad (68)$$

Only if: Suppose \mathbb{J} is not a valid probability distribution. Then there is some $x \in X$ such that $X \bowtie x = \emptyset$ but $\mathbb{J}(x) > 0$. Then

$$\mu^X(x) = \mu(X \bowtie x) \quad (69)$$

$$= \sum_{x' \in X} \mathbb{J}(x') \mathbb{K}(X \bowtie x | x') \quad (70)$$

$$= 0 \quad (71)$$

$$\neq \mathbb{J}(x) \quad (72)$$

□

2.8 Higher order gap models

Order 1 gap models are represented by *conditional probabilities* and order 2 gap models by *probability 2-combs*. Order 1 models take valid candidate distributions and map to order 0 gap models using the “copy-product” \odot . Order 2 models take valid candidate conditionals and, via the “insert” operation, yield over 1 models.

A note on terminology: Probability gap models are written with black-board letters \mathbb{P} . The same base letter with different superscripts $\mathbb{P}^A|B$ indicates

a conditional probability with respect to \mathbb{P} . We use subscripts to indicate application of the model to an insert $\mathbb{P}_\alpha := \mathbb{P}(\mathbb{P}_\alpha^X)$, and \mathbb{P}_α is the resulting model of lower order. The same base letter with different superscripts $\mathbb{P}^{A|B}$ indicates a conditional probability with respect to \mathbb{P} , and $A \perp\!\!\!\perp_{\mathbb{P}} B$ is a statement of independence with respect to the model \mathbb{P} .

Definition 2.17 (Order 1 probability gap model). Given probability set Ω and variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$, denote by A the set of valid candidate distributions on X . Then an order 1 probability gap model \mathbb{P} is a function from $A \rightarrow \Delta(X \times Y)$ such that some $\mathbb{P}^{Y|X} : X \rightarrow Y$ is a conditional probability with respect to \mathbb{P} .

Definition 2.18 (Order 1 model defined by a conditional). Given probability set Ω and variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$, A the set of valid candidate distributions on X and valid candidate conditional $\mathbb{P}^{Y|X}$, define $\mathbb{P} : A \rightarrow \Delta(X \times Y)$ to be the order 1 model such that defined by

$$\mathbb{P}_\alpha^X \mapsto \mathbb{P}_\alpha^X \odot \mathbb{P}^{Y|X} \quad (73)$$

For any valid candidate distribution $\mathbb{P}_\alpha^X \in A$.

The notation $\mathbb{P}^{\square Y|X}$ is shorthand for an order 1 model \mathbb{P} defined by a conditional $\mathbb{P}^{Y|X}$.

Theorem 2.22 shows that Definition 2.18 is indeed an order 1 model.

First, we show that Definitions 2.12 and 2.13 define equivalent notions of validity.

Lemma 2.19 (Equivalence of validity definitions). *Given $X : \Omega \rightarrow X$, with Ω and X discrete, a probability measure $\mathbb{P}^X \in \Delta(X)$ is valid if and only if the conditional $\mathbb{P}^{X|*} := * \mapsto \mathbb{P}^X$ is valid.*

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

$$\forall x \in X : X \bowtie (x) = \emptyset \implies \mathbb{P}^{X|*}(x|*) = 0 \quad (74)$$

$$= \mathbb{P}^X(x) \quad (75)$$

If: We refer to Ershov (1975) Theorem 2.5 for the proof that Equation 75 is necessary and sufficient for the existence of \mathbb{P}^I such that $\mathbb{P}^I(X^{-1}(A)) = \mathbb{P}^X(A)$ for all $A \in \mathcal{X}$ when (Ω, \mathcal{F}) and (X, \mathcal{X}) are standard measurable. If Ω and X are discrete, then they are standard measurable.

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

Next, we show that

Lemma 2.20 (Product of valid candidate conditionals is valid). *Given (Ω, \mathcal{F}) , $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, $Z : \Omega \rightarrow Z$ and any valid candidate conditional $\mathbb{P}^{Y|X}$ and $\mathbb{Q}^{Z|YX}$, $\mathbb{P}^{Y|X} \odot \mathbb{Q}^{Z|YX}$ is also a valid candidate conditional.*

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

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

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

$$\mathbb{R}^{YZ|X}(y, z|x) = \mathbb{Q}^{Z|YX}(z|y, x) \mathbb{P}^{Y|X}(y|x) \quad (76)$$

$$\leq \mathbb{P}^{Y|X}(y|x) \quad (77)$$

$$= 0 \quad (78)$$

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

$$\mathbb{R}^{YZ|X}(y, z|x) = \mathbb{Q}^{Z|YX}(z|y, x) \mathbb{P}^{Y|X}(y|x) \quad (79)$$

$$= 0 \quad (80)$$

□

Corollary 2.21 (Valid candidate conditional is validly extendable to a valid candidate distribution). *Given Ω , $U : \Omega \rightarrow U$, $W : \Omega \rightarrow W$ and a valid candidate conditional $\mathbb{T}^{W|U}$, then for any valid candidate conditional \mathbb{V}^U , $\mathbb{V}^U \odot \mathbb{T}^{W|U}$ is a valid candidate probability.*

Proof. Applying Lemma 2.20 choosing $X = *$, $Y = U$, $Z = W$ and $\mathbb{P}^{Y|X} = \mathbb{V}^{U|*}$ and $\mathbb{Q}^{Z|YX} = \mathbb{T}^{W|U*}$ we have $\mathbb{R}^{WU|*} := \mathbb{V}^{U|*} \odot \mathbb{T}^{W|U*}$ is a valid conditional probability. Then $\mathbb{R}^{WU} \cong \mathbb{R}^{WU|*}$ is valid by Theorem 2.19. □

Theorem 2.22 (Validity of conditional probabilities). *Suppose we have Ω , $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, with Ω , X , Y discrete. A conditional $\mathbb{T}^{Y|X}$ is valid if and only if for all valid candidate distributions \mathbb{V}^X , $\mathbb{V}^X \odot \mathbb{T}^{Y|X}$ is also a valid candidate distribution.*

Proof. If: this follows directly from Corollary 2.21.

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

$$(\mathbb{V}^X \odot \mathbb{T}^{Y|X})(x, y) = \mathbb{T}^{Y|X}(y|x) \mathbb{V}^X(x) \quad (81)$$

$$= \mathbb{T}^{Y|X}(y|x) \quad (82)$$

$$> 0 \quad (83)$$

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

2.9 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 represented by *probability 2-combs* (Chiribella et al., 2008; Jacobs et al., 2019).

Definition 2.23 (Probability 2-comb). Given $W : \Omega \rightarrow W$, $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, $Z : \Omega \rightarrow Z$, a probability 2-comb $\mathbb{P}^{X|W \square Z|Y} : W \times Y \rightarrow X \times Z$ is a Markov kernel such that for some $\mathbb{P}^{X|W} : W \rightarrow X$

$$\begin{array}{c} W \\ Y \end{array} \begin{array}{c} \boxed{\mathbb{P}^{X|W \square Z|Y}} \\ \hline \end{array} \begin{array}{c} X \\ * \end{array} = \begin{array}{c} W \\ Y \end{array} \begin{array}{c} \boxed{\mathbb{P}^{X|W}} \\ \hline \end{array} \begin{array}{c} X \\ * \end{array} \quad (84)$$

Definition 2.24 (Valid probability 2-comb). $\mathbb{P}^{X|W \square Z|Y} : W \times Y \rightarrow X \times Z$ is a valid probability 2-comb if

1. $\mathbb{P}^{X|W}$ is a valid 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 $\mathbb{P}^{Z|WXY} : W \times X \times Y \rightarrow Z$ (Lemma 2.29) such that

$$\mathbb{P}^{X|W \square Z|Y} = \begin{array}{c} W \\ Y \end{array} \begin{array}{c} \bullet \\ \bullet \end{array} \begin{array}{c} \boxed{\mathbb{P}^{X|W}} \\ \hline \end{array} \begin{array}{c} X \\ \bullet \end{array} \begin{array}{c} \boxed{\mathbb{P}^{Z|WXY}} \\ \hline \end{array} \begin{array}{c} Z \\ \bullet \end{array} \quad (85)$$

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

$$\text{insert}(\mathbb{P}_\alpha^{Y|XW}, \mathbb{P}^{X|W \square Z|Y}) = \mathbb{P}^{X|W} \odot \mathbb{P}_\alpha^{Y|XW} \odot \mathbb{P}^{Z|WXY} \quad (86)$$

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

$$\text{Insert} \left(\begin{array}{c} \begin{array}{c} W \\ \bullet \end{array} \begin{array}{c} \boxed{\mathbb{P}^{X|W}} \\ \hline \end{array} \begin{array}{c} X \\ \bullet \end{array} \begin{array}{c} \boxed{\mathbb{P}^{Z|XYW}} \\ \hline \end{array} \begin{array}{c} Z \\ \bullet \end{array} \\ , \\ \begin{array}{c} X \\ W \end{array} \begin{array}{c} \boxed{\mathbb{P}_\alpha^{Y|XW}} \\ \hline \end{array} \begin{array}{c} X \\ \bullet \end{array} \end{array} \right) \quad (87)$$

$$= \begin{array}{c} W \\ \bullet \end{array} \begin{array}{c} \boxed{\mathbb{P}^{X|W}} \\ \hline \end{array} \begin{array}{c} \bullet \end{array} \begin{array}{c} \boxed{\mathbb{P}_\alpha^{Y|XW}} \\ \hline \end{array} \begin{array}{c} \bullet \end{array} \begin{array}{c} \boxed{\mathbb{P}^{Z|WXY}} \\ \hline \end{array} \begin{array}{c} Z \\ \bullet \end{array} \begin{array}{c} Y \\ \bullet \end{array} \begin{array}{c} X \\ \bullet \end{array} \quad (88)$$

Definition 2.25 (Order 2 probability gap model). An order 1 probability gap model \mathbb{P} is defined by variables $W : \Omega \rightarrow W$, $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ and $X : \Omega \rightarrow Z$ and a valid conditional probability 2-comb $\mathbb{P}^{X|W \square Z|Y}$, and \mathbb{P} is the

map from valid $Y|X$ conditional probabilities to valid $(X, Y, Z)|W$ conditional probabilities given by

$$\mathbb{P}_\alpha^{XYZ|W} = \mathbb{P}^{X|W} \odot \mathbb{P}_\alpha^{Y|X} \odot \mathbb{P}^{Z|WXY} \quad (89)$$

for any valid $Y|X$ conditional probability $\mathbb{P}_\alpha^{Y|X}$.

Definition 2.26 (Order 2 model defined by a 2-comb).

An order 2 probability gap model $\mathbb{P}^{X|W \square Z|Y}$ can be defined by a pair of disintegrations $\mathbb{P}^{X|W}$ and $\mathbb{P}^{Z|WXY}$. Just as with order 1 probability gap models, the two conditional probabilities defining a probability 2-comb must be *valid*. Theorem 2.27 shows that the insert operation yields a valid conditional probability given any valid inputs.

Theorem 2.27 (Extension of valid probability 2-combs). *Given Ω , $W : \Omega \rightarrow W$, $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ and $Z : \Omega \rightarrow Z$, a probability 2-comb $\mathbb{P}^{X|W \square Z|Y}$ is valid if and only if $\text{insert}(\mathbb{P}_\alpha^{Y|WX}, \mathbb{P}^{X|W \square Z|Y})$ is valid for all valid $\mathbb{P}_\alpha^{Y|WX}$.*

Proof. Only if:

Note that

$$\mathbb{P}_\alpha^{XYZ|W} := \text{insert}(\mathbb{P}_\alpha^{Y|WX}, \mathbb{P}^{X|W \square Z|Y}) \quad (90)$$

$$\mathbb{P}_\alpha^{XYZ|W}(xyz|w) = \mathbb{P}_\alpha^{Y|WX}(y|w, x) \mathbb{P}^{X|W}(x|w) \mathbb{P}^{Z|XYW}(z|x, y, w) \quad (91)$$

$$= \mathbb{P}^{X|W \square Z|Y}(x, z|w, y) \mathbb{P}_\alpha(y|x) \quad (92)$$

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

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

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

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

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

$$\mathbb{P}_\alpha^{X|W}(x|w) = \sum_{y \in Y, z \in Z} \mathbb{P}_\alpha^{XYZ|W}(xyz|w) \quad (93)$$

$$= \mathbb{P}^{X|W}(x|w) \quad (94)$$

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

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

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

$$\mathbb{P}_\alpha^{XYZ|W}(xyz|w) = \mathbb{P}^{X|W \square Z|Y}(x, z|w, y) \mathbb{P}_\alpha(y|x) \quad (95)$$

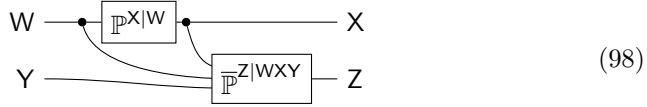
$$= \mathbb{P}^{X|W \square Z|Y}(x, z|w, y) \quad (96)$$

$$> 0 \quad (97)$$

So $\mathbb{P}_\alpha^{XYZ|W}$ is invalid. \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 2.28 (Valid conditional probabilities combine to a valid 2-comb).
Given valid conditional probabilities $\mathbb{P}^{X|W}$ and $\mathbb{P}^{Z|WXY}$



Is a valid 2-comb.

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

2.10 Revisiting truncated factorisation

In our original look at truncated factorisation, we noted a few problems with Equation 42 being a *definition* of interventional probability models. In particular:

- There may be multiple different probability models that satisfy Equation 42 for different versions of the disintegration $\mathbb{P}^{Y|XZ}$
- There may be no probability models that satisfy Equation 42

We propose a different way to define interventional probability models

- An interventional probability model is a probability 2-comb $\mathbb{Q}^{Z \square Y|X}$
- Observations are distributed according to $\mathbb{Q}_{\text{obs}} := \text{insert}(\mathbb{Q}_{\text{obs}}^{X|Z}, \mathbb{Q}^{Z \square Y|X})$

In this case, we have

$$\mathbb{Q}_{\text{obs}}^Z = \mathbb{Q}^Z \quad (99)$$

$$\mathbb{Q}^{Y|XZ} \subset \mathbb{Q}_{\text{obs}}^{Y|XZ} \quad (100)$$

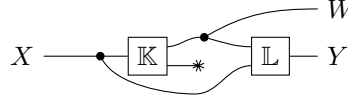
Now, for $x \in X$ if the deterministic insert $\mathbb{Q}_x^{X|Z}$ defined by $\mathbb{Q}_x^{X|Z}(x'|z) = \llbracket x = x' \rrbracket$ is a valid conditional probability, then for some version of $\mathbb{Q}^{Y|XZ}$ the following holds:

$$\mathbb{T}_x^{XYZ} = \mathbb{Q}_{\text{obs}}^{Y|XZ}(y|x, z) \mathbb{Q}_{\text{obs}}^Z(z) \llbracket x = x' \rrbracket \quad (101)$$

This is identical in form to Equation 42, but we have made explicit two assumptions that were implicit in that equation - namely, the validity of hard interventions on X and the possibility of appropriately choosing a version of $\mathbb{Q}^{Y|XZ}$.

2.10.1 Disintegrations

Lemma 2.29 (Disintegration existence in discrete Markov kernels). *For any Markov kernel $\mathbb{K} : X \rightarrow W \times Y$ and X, W, Y are discrete, there exists $\mathbb{L} : W \times X \rightarrow Y$ such that*



$$\mathbb{K} = \quad (102)$$

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

$$\mathbb{L}(y|w, x) = \frac{\mathbb{K}(w, y|x)}{\sum_{y \in Y} \mathbb{K}(w, y|x)} \quad \forall x, w : \text{the denominator is positive} \quad (103)$$

Then

$$\sum_{y \in Y} \mathbb{K}(w, y|x) \mathbb{L}(y|w, x) = \sum_{y \in Y} \mathbb{K}(w, y|x) \frac{\mathbb{K}(w, y|x)}{\sum_{y \in Y} \mathbb{K}(w, y|x)} \quad \text{if } \sum_{y \in Y} \mathbb{K}(w, y|x) > 0 \quad (104)$$

$$= \mathbb{K}(w, y|x) \quad \text{if } \sum_{y \in Y} \mathbb{K}(w, y|x) > 0 \quad (105)$$

$$= 0 \quad \text{otherwise} \quad (106)$$

$$= \mathbb{K}(w, y|x) \quad \text{otherwise} \quad (107)$$

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

Theorem 2.30 (Existence of conditional probabilities). *Given valid $\mathbb{K}^{\square WY|X}$, there exists a valid conditional probability $\mathbb{K}^{Y|WX}$.*

Proof. From Lemma 2.29, we have the existence of some Markov kernel $\mathbb{K}^{Y|WX} : W \times X \rightarrow Y$ such that

$$\mathbb{K}^{WY|X} = \mathbb{K}^{W|X} \odot \mathbb{K}^{Y|WX} \quad (108)$$

We need to check that $\mathbb{K}^{Y|WX}$ 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 $\mathbb{K}^{W,Y|X}(w, y|x) = 0 \implies \mathbb{K}^{Y|WX}(y|w, x) = 0$. 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 $\mathbb{K}^{Y|WX}(Y(\omega)|w, x) = 1$.

Given that $\mathbb{K}^{Y|WX}$ is valid, we also require that for all completions \mathbb{K}_α^{XWY} of $\mathbb{K}^{\square WY|X}$,

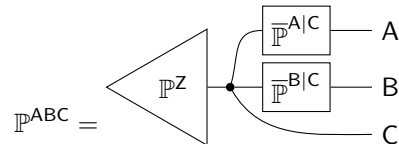
$$\mathbb{K}_\alpha^{XWY} = \mathbb{K}_\alpha^{XW} \odot \mathbb{K}^{Y|WX} \quad (109)$$

Noting that $\mathbb{K}_\alpha^{XWY} = \mathbb{K}^{W|X} \odot \mathbb{K}^{Y|WX}$, this follows directly from Equation 108. \square

2.10.2 Conditional independence

Conditional independence has a familiar definition in probability models. A conditional independence in models with probability gaps is equivalent to the claim that the given conditional independence holds for all base measure that the model can be extended to. In order 1 probability gap models, this definition is closely related to the idea of *extended conditional independence* proposed by Constantinou and Dawid (2017).

Definition 2.31 (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



$$\mathbb{P}^{ABC} = \quad (110)$$

For any $\mathbb{P}^{B|C}$ and $\mathbb{P}^{A|C}$. 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}^W if it holds for all probability models \mathbb{P}_{α}^I that pushforward to \mathbb{P}^W . This might happen when, for example, $(A, B, C) = W$. Thus we define order 0 conditional independence with respect to an arbitrary probability distribution as well.

Theorem 2.32.

Equivalence of different statements of conditional independence

Definition 2.33 (Order 1 conditional independence). For a *conditional probability model* $\mathbb{P}^{V|W}$, define $\mathbb{P}_{\beta}^{VW} := \mathbb{P}_{\beta}^W \odot \mathbb{P}^{V|W}$. Then say B is order 1 conditionally independent of A given C , written $B \perp\!\!\!\perp_{\mathbb{P}}^1 A|C$ if for all β , $B \perp\!\!\!\perp_{\mathbb{P}_{\beta}} A|C$.

In the case of a conditional probability model $\mathbb{P}^{V|W}$, we can view $\mathbb{P}^{V|W}$ as a collection of probability distributions $\{\mathbb{P}_w^V := A \mapsto \mathbb{P}^{V|W}(A|w) | w \in W\}$. 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 $\mathbb{P}^{V|W}$. Theorem 4.4 of that work proves the following claim:

Theorem 2.34. 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}_{α}^W be some probability distribution on W such that $w \in W(\Omega) \implies \mathbb{P}_{\alpha}^W(w) > 0$. Then, denoting extended conditional independence with $\perp\!\!\!\perp_{\mathbb{P}, ext}$ and $\mathbb{P}_{\alpha}^{VW} := \mathbb{P}_{\alpha}^W \odot \mathbb{P}^{V|W}$

$$A \perp\!\!\!\perp_{\mathbb{P}, ext} (B, D)|(C, E) \iff A^* \perp\!\!\!\perp_{\mathbb{P}_{\alpha}} (B^*, D^*|(C^*, E^*) \quad (111)$$

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

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

Theorem 2.35. 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 $\mathbb{P}_{\alpha}^{VW} := \mathbb{P}_{\alpha}^W \odot \mathbb{P}^{V|W}$

$$A \perp\!\!\!\perp_{\mathbb{P}, ext}^1 (B, D)|(C, E) \iff A^* \perp\!\!\!\perp_{\mathbb{P}} (B^*, D^*|(C^*, E^*) \quad (112)$$

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 2.34, $A \perp\!\!\!\perp_{\mathbb{P}, ext} (B, D)|(C, E)$.

Only if:

For any β , $\mathbb{P}_{\beta}^{ABC|DE} = \mathbb{P}_{\beta}^{DE} \odot \mathbb{P}^{ABC|DE}$. By Lemma 2.29, we have $\mathbb{P}^{A|BCDE}$ such that

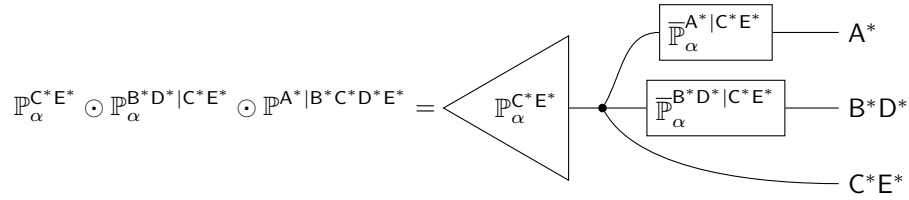
$$\mathbb{P}_\beta^{A^*B^*C^*D^*E^*} = \mathbb{P}_\beta^{D^*E^*} \odot \mathbb{P}^{B^*C^*|D^*E^*} \odot \mathbb{P}^{A^*|B^*C^*D^*E^*} \quad (113)$$

$$= \mathbb{P}_\beta^{B^*C^*D^*E^*} \odot \mathbb{P}^{A^*|B^*C^*D^*E^*} \quad (114)$$

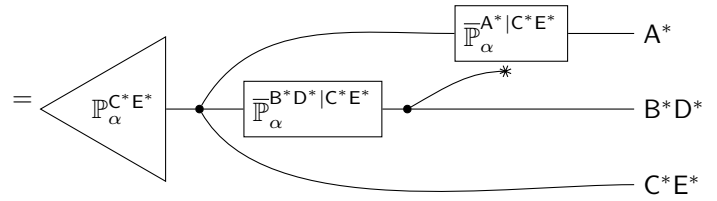
$$= \mathbb{P}_\beta^{C^*E^*} \odot \mathbb{P}_\beta^{B^*D^*|C^*E^*} \odot \mathbb{P}^{A^*|B^*C^*D^*E^*} \quad (115)$$

By Theorem 2.34, we have some α such that $\mathbb{P}_\alpha^{D^*E^*}$ is strictly positive on the range of (D^*, E^*) and $A^* \perp_{\mathbb{P}_\alpha} (B^*, D^*)|(C^*, E^*)$.

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



(116)



(117)

$$= \mathbb{P}_\alpha^{C^*E^*} \odot \mathbb{P}_\alpha^{B^*D^*|C^*E^*} \odot (\mathbb{P}_\alpha^{A^*|C^*E^*} \otimes \text{erase}_{BD}) \quad (118)$$

Thus for any $(a, b, c, d, e) \in A \times B \times C \times D \times E$ such that $\mathbb{P}_\alpha^{B^*C^*D^*E^*}(b, c, d, e) > 0$, $\mathbb{P}^{A^*|B^*C^*D^*E^*}(a|b, c, d, e) = \mathbb{P}_\alpha^{A^*|C^*E^*}(a|c, e)$ 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 $\mathbb{P}_\beta^{A^*|B^*C^*D^*E^*} \cong \mathbb{P}_\alpha^{A^*|C^*E^*}(a|c, e)$ also. But then

$$\begin{aligned}
\mathbb{P}_\beta^{A^*B^*C^*D^*E^*} &= \triangleleft \mathbb{P}_\beta^{C^*E^*} \begin{array}{c} \text{---} \bullet \text{---} \begin{array}{c} \boxed{\mathbb{P}_\alpha^{A^*|C^*E^*}} \text{---} A^* \\ \boxed{\mathbb{P}_\beta^{B^*D^*|C^*E^*}} \text{---} B^*D^* \\ \text{---} C^*E^* \end{array} \end{array} \\
&\quad \quad \quad (119)
\end{aligned}$$

$$\begin{aligned}
&= \triangleleft \mathbb{P}_\beta^{C^*E^*} \begin{array}{c} \boxed{\mathbb{P}_\alpha^{A^*|C^*E^*}} \text{---} A^* \\ \boxed{\mathbb{P}_\beta^{B^*D^*|C^*E^*}} \text{---} B^*D^* \\ \text{---} C^*E^* \end{array} \\
&\quad \quad \quad (120)
\end{aligned}$$

□

Definition 2.36 (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}_α^I , \mathbb{P}_β^{VW} and $\mathbb{P}_\gamma^{XYZ|W}$, and so they also hold for the higher level definition of conditional probability.

2.11 Recursive disintegration

A useful property of probability gap models is *recursive disintegration*. If we have any probability gap model \mathbb{P} and a disintegration $\mathbb{P}^{Y|X}$, then defining a new order 1 probability gap model $\mathbb{Q}^{\square Y|X}$ such that $\mathbb{Q}^{Y|X} = \mathbb{P}^{Y|X}$, disintegrations of \mathbb{Q} are themselves disintegrations of \mathbb{P} .

Corollary 2.37 (Recursive independence). *Suppose we have a fundamental probability set Ω , a set of inserts A , variables $V : \Omega \rightarrow V$, $W : \Omega \rightarrow W$, $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, $Z : \Omega \rightarrow Z$ such that $Y = (Z, W)$ and a probability gap model $\mathbb{P} : A \rightarrow \Delta(\Omega)$ such that $\mathbb{P}^{X|Y}$ is a $Y|X$ disintegration of \mathbb{P} . Define an order 1 probability gap model $\mathbb{Q}^{\square Y|X}$ such that $\mathbb{Q}^{Y|X} = \mathbb{P}^{Y|X}$. If $Z \perp\!\!\!\perp_{\mathbb{Q}} W|V$ then $Z \perp\!\!\!\perp_{\mathbb{P}} W|V$.*

Proof. If $Z \perp\!\!\!\perp_{\mathbb{Q}} W|V$ then both $\mathbb{Q}^{Z|WV}$ and $\mathbb{Q}^{Z|V}$ exist and $\mathbb{Q}^{Z|WV} = \mathbb{Q}^{Z|V} \otimes \text{erase}_W$ and hence $\mathbb{P}^{Z|WV} = \mathbb{P}^{Z|V} \otimes \text{erase}_W$. □

2.11.1 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 2.38. 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 2.38. *Given a probability 2-comb $\mathbb{P}^{XV|W \square Y|XV}$, $Y \perp\!\!\!\perp_{\mathbb{P}}^2 V|WX$ if and only if there is a version of $\mathbb{P}^{Y|WXV}$ and some $\mathbb{P}^{Y|WX}$ such that*

$$\mathbb{P}^{Y|WXV} = \begin{array}{c} W \\ X \\ V \end{array} \begin{array}{c} \diagup \\ \diagdown \\ \longrightarrow \end{array} \boxed{\mathbb{P}^{Y|WX}} \begin{array}{c} \diagdown \\ \diagup \\ \longrightarrow \end{array} Y \quad (121)$$

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

$$\mathbb{L} := \mathbb{J} \odot \text{insert}(\mathbb{P}^{XV|W \square Y|XV}, \mathbb{K}) \quad (122)$$

$$= \mathbb{J} \odot \mathbb{P}^{XV|W} \odot \mathbb{K} \odot \mathbb{P}^{Y|WXV} \quad (123)$$

$$(124)$$

That is, $\mathbb{P}^{Y|WXV}$ is a version of $\mathbb{L}^{Y|WXV}$ for any extension \mathbb{L} of \mathbb{P} . Then by Theorem 2.32, $Y \perp\!\!\!\perp_{\mathbb{L}} V|WX$ and so, by definition, $Y \perp\!\!\!\perp_{\mathbb{P}}^2 V|WX$. Only if:

Let $\mathbb{L} : \{*\} \rightarrow W \times X \times V \times Y$ be an extension of \mathbb{P} to a 0-order model such that $\mathbb{L}^{WXV} \gg \mathbb{M}^{WXV}$ for any other extension \mathbb{M} . Because $Y \perp\!\!\!\perp_{\mathbb{L}} V|WX$, by Theorem 2.32 there is some $\mathbb{L}^{Y|WXV}$ and $\mathbb{L}^{Y|WX}$ such that

$$\mathbb{L}^{Y|WXV} = \begin{array}{c} W \\ X \\ V \end{array} \begin{array}{c} \diagup \\ \diagdown \\ \longrightarrow \end{array} \boxed{\mathbb{L}^{Y|WX}} \begin{array}{c} \diagdown \\ \diagup \\ \longrightarrow \end{array} Y \quad (125)$$

As \mathbb{L} is an extension of \mathbb{P} , there must be some \mathbb{J}, \mathbb{K} such that for any $\mathbb{P}^{Y|WXV}$

$$\mathbb{L} := \mathbb{J} \odot \text{insert}(\mathbb{P}^{XV|W \square Y|XV}, \mathbb{K}) \quad (126)$$

$$= \mathbb{J} \odot \mathbb{P}^{XV|W} \odot \mathbb{K} \odot \mathbb{P}^{Y|WXV} \quad (127)$$

$$\mathbb{L}^{WXV} = \mathbb{J} \odot \mathbb{P}^{XV|W} \odot \mathbb{K} \quad (128)$$

Thus for any (w, x, v) such that $\mathbb{L}^{WXV} > 0$, $\mathbb{P}^{Y|WXV} \stackrel{\mathbb{L}}{\cong} \mathbb{L}^{Y|WXV}$. The set on which they may disagree is precisely the set of (w, x, v) such that $\mathbb{J} \odot \mathbb{P}^{XV|W} \odot \mathbb{K} = 0$ for all \mathbb{J}, \mathbb{K} , but this means that $\mathbb{P}^{Y|WXV} \stackrel{\mathbb{P}}{\cong} \mathbb{L}^{Y|WXV}$ also. \square

2.11.2 Restricted 2-combs

this notation sucks

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 $\mathbb{P}^{X|W} \boxed{A}^{Y|D}$.

Given $\mathbb{P}^{X|W \square Y|D}$ and some random variable $V = f \circ (W, X)$, define the subset of inserts that depend only on V as those $\mathbb{Q}_\alpha^{D|XW}$ such that $D \perp\!\!\!\perp_{\mathbb{Q}}^1 (W, X) | V$. Define B as the set of all inserts exhibiting this conditional independence. Then write $\mathbb{P}^{X|W} \boxed{V}^{Y|D}$ for the restriction of $\mathbb{P}^{X|W \square Y|D}$ 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 $\mathbb{P}^{X|W} \boxed{*}^{Y|D}$. For any (X, W) , $* = \text{erase}_{X \times W} \circ (X, W)$, so this is a well-defined restriction. However, we must still assume that there exists at least one $\mathbb{Q}_\alpha^{D|XW}$ such that $D \perp\!\!\!\perp_{\mathbb{Q}}^1 (W, X) | *$, 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 2.39 (Restricted order 2 conditional independence). For a *restricted probability 2-comb* $\mathbb{P}^{X|W} \boxed{E}^{Z|Y}$, define $\mathbb{P}_\gamma^{XYZ|W} := \text{insert}(\mathbb{Q}_\gamma^{Y|XW}, \mathbb{P}^{X|W \square Z|Y})$ for any $\mathbb{Q}_\gamma \in E$. Then B is restricted order 2 conditionally independent of A given C , written $B \perp\!\!\!\perp_{\mathbb{P}}^{2|E} A | C$ if for all $\gamma \in E$, $B \perp\!\!\!\perp_{\mathbb{P}_\gamma}^1 A | C$.

2.12 Results I use that don't really fit into the flow of the text

2.12.1 Repeated variables

Lemmas 2.40 and 2.41 establish that models of repeated variables must connect the repetitions with a copy map.

Lemma 2.40 (Output copies of the same variable are identical). *For any Ω , X, Y, Z random variables on Ω and conditional probability $\mathbb{K}^{YZ|X}$, there is a conditional probability $\mathbb{K}^{YYZ|X}$ unique up to impossible values of X such that*

$$X \text{ --- } \boxed{\mathbb{K}^{YYZ|X}} \begin{array}{c} \text{---}^* \\ \text{---} \\ \text{---} \end{array} Y \text{ --- } Z = \mathbb{K}^{YZ|X} \quad (129)$$

and it is given by

$$\mathbb{K}^{\mathbf{Y}\mathbf{Y}\mathbf{Z}|\mathbf{X}} = \mathbf{X} \text{ --- } \boxed{\mathbb{K}^{\mathbf{Y}\mathbf{Z}|\mathbf{X}}} \begin{array}{l} \text{--- } \mathbf{Y} \\ \text{--- } \mathbf{Y} \\ \text{--- } \mathbf{Z} \end{array} \quad (130)$$

$$\iff \quad (131)$$

$$\mathbb{K}^{\mathbf{Y}\mathbf{Y}\mathbf{Z}|\mathbf{X}}(y, y', z|x) = \llbracket y = y' \rrbracket \mathbb{K}^{\mathbf{Y}\mathbf{Z}|\mathbf{X}}(y, z|x) \quad (132)$$

$$(133)$$

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

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

$$\mathbb{K}^{\mathbf{Y}\mathbf{Y}\mathbf{Z}|\mathbf{X}}(y, y', z|x) = \llbracket y = y' \rrbracket \mathbb{L}(y, z|x) \quad (134)$$

$$(135)$$

But then

$$\mathbb{K}^{\mathbf{Y}\mathbf{Z}|\mathbf{X}}(y, z|x) = \sum_{y' \in \mathbf{Y}} \mathbb{K}^{\mathbf{Y}\mathbf{Y}\mathbf{Z}|\mathbf{X}}(y, y', z|x) \quad (136)$$

$$= \mathbb{L}(y, z|x) \quad (137)$$

$$(138)$$

□

Lemma 2.41 (Copies shared between input and output are identical).

This got mixed up at some point and needs ot be unmixed-up

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

$$\mathbb{K} = \begin{array}{c} \mathbf{X} \\ \mathbf{Y} \end{array} \text{ --- } \boxed{\mathbb{K}^{\mathbf{Z}|\mathbf{X}\mathbf{Y}}} \begin{array}{l} \text{--- } \mathbf{X} \\ \text{--- } \mathbf{Z} \end{array} \quad (139)$$

$$\iff \quad (140)$$

$$\mathbb{K}_{x,y}^{\mathbf{I}|\mathbf{x}',z} = \llbracket x = x' \rrbracket \mathbb{L}_{x,y}^z \quad (141)$$

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

$$\begin{array}{c} \mathbf{X} \\ \mathbf{Y} \end{array} \text{ --- } \boxed{\mathbb{K}^{\mathbf{X}\mathbf{Z}|\mathbf{X}\mathbf{Y}}} \begin{array}{l} \text{--- } * \\ \text{--- } \mathbf{Z} \end{array} = \mathbb{K}^{\mathbf{X}\mathbf{Z}|\mathbf{X}\mathbf{Y}} \quad (142)$$

and it is given by

$$\mathbb{K}^{XZ|XY} = X \text{ --- } \boxed{\mathbb{K}^{YZ|X}} \begin{array}{l} \text{--- } Y \\ \text{--- } Y \\ \text{--- } Z \end{array} \quad (143)$$

$$\iff \quad (144)$$

$$\mathbb{K}^{XZ|XY}(x, z|x', y) = \llbracket x = x' \rrbracket \mathbb{K}^{Z|XY}(z|x', y) \quad (145)$$

$$(146)$$

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

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

$$\mathbb{K}^{XZ|XY}(x', z|x, y) = 0 = \llbracket x = x' \rrbracket L(z|x, y) \quad (147)$$

$$(148)$$

But then

$$\mathbb{K}^{Z|XY}(y, z|x) = \sum_{x' \in X} \mathbb{K}^{XZ|XY}(x', z|x, y) \quad (149)$$

$$= L(z|x, y) \quad (150)$$

$$(151)$$

□

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 \mathcal{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 \mathcal{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* $\mathbb{Q}_\alpha^{D|X} : X \rightarrow D$. That is, we need a probability gap model \mathbb{P} that takes a decision function $\mathbb{Q}_\alpha^{X|D}$ and returns a probabilistic model of the consequences of selecting that decision function $\mathbb{P}_\alpha^{DXY} := M(\mathbb{Q}_\alpha^{D|X})$. An order 2 model $\mathbb{P}^{X \square Y | D}$ is such a function, but there are many such functions that are not probability 2-combs. We will define *ordinary decision problems* to be those for which the desired model \mathbb{P} is an order 2 probability gap model.

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. To do this, we include an unobserved variable H , the *hypothesis*, which can be interpreted as representing another gap in our knowledge: if we knew the value of H , or 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 $\mathbb{P}^{X | H \square Y | D}$.

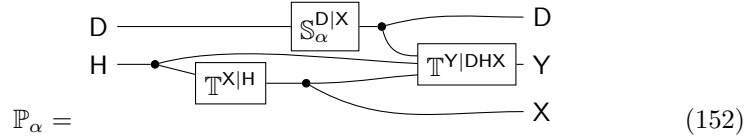
Specifically, an *ordinary decision problem* M features hypotheses $H : \Omega \rightarrow H$,

observations $X : \Omega \rightarrow X$, decisions $D : \Omega \rightarrow D$ and consequences $Y : \Omega \rightarrow Y$. It is modeled by a function $M : \Delta(D)^X \rightarrow \Delta(X \times D \times Y)^H$ such that, defining $\mathbb{P}_\alpha^{DXY} := M(Q_\alpha^{D|X})$,

1. There is some $\mathbb{P}_\alpha^{D|X}$ such that $Q_\alpha^{D|X} = \mathbb{P}_\alpha^{D|X}$ for all decision functions α
2. There is some $\mathbb{P}^{X|H}$ such that, for all decision functions α , there is some $\mathbb{P}_\alpha^{X|H}$ such that $\mathbb{P}^{X|H} = \mathbb{P}_\alpha^{X|H}$
3. There is some $\mathbb{P}^{Y|XDH}$ such that, for all decision functions α , there is some $\mathbb{P}_\alpha^{Y|XDH}$ such that $\mathbb{P}^{Y|XDH} = \mathbb{P}_\alpha^{Y|XDH}$

(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 :



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.

3.3 Unresponsiveness

Given a see-do model $\mathbb{P}^{X|H \square Y|D}$, $A \perp\!\!\!\perp_{\mathbb{P}}^{2|*} D|H$ 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 $A \in \{0, 1\}$ representing the outcome of a fair coin toss, $D \in \{0, 1\}$ representing a bet on the coin toss and $B \in \{0, 1\}$ 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 $A \perp\!\!\!\perp_{\mathbb{P}}^{2|*} D|H$ (“the coin toss is independent of the decision, for any decision rule”) and $B \perp\!\!\!\perp_{\mathbb{P}}^{2|*} D|H$ (“the outcome of the bet is independent of the decision, for any decision rule”). However, it is not the case that $(A, B) \perp\!\!\!\perp_{\mathbb{P}}^{2|*} D|H$. 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 $A \perp\!\!\!\perp_{\mathbb{P}}^{2|*} D|H$; specifically, where a variable is independent of D given H under the restricted 2-comb $\mathbb{P}^{X|H \square Y|D}$.

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 measurement processes carried out during observations and after taking action. This can be thought of as the causal analogue of the difference between saying we have a sequence (X_1, X_2, X_3) of observations independent and identically distributed according to $\mu \in \Delta(X)$ and saying that we have some observations distributed according to $\mathbb{P}^X \in \Delta(X)$. 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 $\mathbb{P}^{X|H \square Y|D}$ is $Y \perp\!\!\!\perp_{\mathbb{P}}^2 X|(H, D)$. 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 $\mathbb{P}^{X|H \square Y|D}$ along with the principle of expected utility to evaluate strategies, and the assumption $Y \perp\!\!\!\perp_{\mathbb{P}}^2 X|(H, D)$ 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 $X : \Omega \rightarrow X$ and a statistical experiment $\{\mathbb{P}_\theta^X\}_\Theta$, a decision set D and a loss $l : \Theta \times D \rightarrow \mathbb{R}$. A strategy $S_\alpha^{D|X}$ is evaluated according to the risk functional $R(\theta, \alpha) := \sum_{x \in X} \sum_{d \in D} \mathbb{P}_\theta^X(x) S_\alpha^{D|X}(d|x) l(h, d)$. A strategy $S_\alpha^{D|X}$ is considered more desirable than $S_\beta^{D|X}$ if $R(\theta, \alpha) < R(\theta, \beta)$.

Suppose we have a see-do model $\mathbb{P}^{X|H \square Y|D}$ with $Y \perp\!\!\!\perp_{\mathbb{P}} X|(H, D)$, 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 $l : H \times D \rightarrow \mathbb{R}$ by $l(h, d) = \sum_{y \in \mathbb{R}} y \mathbb{P}^{Y|HD}(y|h, d)$, we have

$$\mathbb{E}_{\mathbb{P}_\alpha}[Y|h] = \sum_{x \in X} \sum_{d \in D} \sum_{y \in Y} \mathbb{P}^{X|H}(x|h) Q_\alpha^{D|X}(d|x) \mathbb{P}^{Y|HD}(y|h, d) \quad (153)$$

$$= \sum_{x \in X} \sum_{d \in D} \mathbb{P}^{X|H}(x|h) Q_\alpha^{D|X}(d|x) l(h, d) \quad (154)$$

$$= R(h, \alpha) \quad (155)$$

If we are given a see-do model where we interpret $\{\mathbb{P}^{X|H}(\cdot|h)|h \in H\}$ as a statistical experiment and Y as a negative utility, the expectation of the utility under the strategy forecast given in equation 152 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 for example there's no precise way in which decisions correspond to interventions. The correspondence exists in the territory, the world that is inhabited by measurement processes, not the 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.

To compare CBNs to see-do models, we will argue that CBNs can be understood as describing probabilistic models of observations and consequences, just like see-do models. Furthermore, CBNs feature an order-1 probability gap and so they describe a probability 2-comb over observations, interventions and consequences. If we suppose that there is some variable describing decisions that does not appear within the CBN, then we can posit a see-do model over observations, decisions and consequences. Finally, we ask: when is the see-do model compatible with the CBN 2-comb, or more precisely, when can we identify each *decision rule* with a *intervention rule* such that the probability model obtained by inserting a decision rule into the see-do model is identical to the probability model obtained by inserting an intervention rule into the CBN 2-comb. We show that see-do models that exhibit a particular type of symmetry are compatible with CBN 2-combs.

4.1 Probability 2-combs represented by causal Bayesian networks

Consider a simplified kind of CBN where a single variable may be intervened on. Note that the structure of the previous section – $X \longrightarrow Y$ and $X \longleftarrow W \longrightarrow Y$ – is generically applicable to such a model if we identify W with the variable formed by taking a sequence of all of the ancestors of X and Y with the variable formed by taking a sequence of all non-ancestors of X . The existence of an edge from X to Y in such a case does no harm as if Y is not “actually” a descendent of X then it will be independent conditional on Z (see Pearl and Mackenzie (2018) for a detailed treatment of when two graphs may or may not imply the same underlying model).

We will adopt the “definitional inversion” discussed in Section 2.10, where we take the causal Bayesian network to express the assumption that the result of intervention is modeled by the order 2 model $\mathbb{Q}^{W \square Y | X}$ and the observations are distributed according to $\text{insert}(\mathbb{Q}_{\text{obs}}^{X|Z}, \mathbb{Q}^{W \square Y | X})$ for some $\mathbb{Q}_{\text{obs}}^{X|Z}$.

We also want to define the fundamental probability set and the variables that we are modelling. We have a sequence of “observation” variables $V_{[n]} := (W_i, X_i, Y_i)_{i \in [n]}$ and a sequence of “consequence” variables modeled by $V_{(n,m]} := (W_i, X_i, Y_i)_{i \in (n,m]}$ both defined on the probability set Ω (assume $n < m$). We suppose that a causal Bayesian network defines a probability gap model \mathbb{Q} of some type for these variables. Suppose for all i , $W_i : \Omega \rightarrow W$, $X_i : \Omega \rightarrow X$, $Y_i : \Omega \rightarrow Y$.

Beyond the choice of intervention, when we have a causal Bayesian network

there is a second gap in our knowledge – in particular, we do not know what $\mathbb{Q}_{\text{obs}}^{X|Z}$ and $\mathbb{Q}^{W \square Y|X}$ are. We address this ignorance of the “correct” probability model as we did in Section 3 by introducing an unobserved hypothesis H . We further assume that, conditional on H , observations are mutually independent; that is, $V_i \perp\!\!\!\perp_{\mathbb{Q}} V_{[n] \setminus \{i\}} | H$ for $i \in [n]$. Each V_i $i \in [n]$ is assumed to be distributed according to $\mathbb{Q}_{\text{obs}}^{W_1 X_1 Y_1 | H}$ (we add the subscripts to avoid ambiguity over what the subscript-less variables refer to). Furthermore, with subscripts in place, the assumption relating observations to consequences reads

$$\mathbb{Q}_{\text{obs}}^{W_1 X_1 Y_1 | H} = \text{insert}(\mathbb{Q}_{\text{obs}}^{X_{n+1} | HW_{n+1}}, \mathbb{Q}^{W_{n+1} \square Y_{n+1} | X_{n+1}}) \quad (156)$$

We make similar “mutually independent and identically distributed” assumptions for consequences, but they are a bit more subtle to state. We assume that, conditional on H , W_j are mutually independent for all $j \in [m]$ and conditional on H and X_j , Y_j are mutually independent for all $j \in [m]$. That is, $W_j \perp\!\!\!\perp_{\mathbb{Q}} V_{[m] \setminus \{j\}} | H$ and $Y_j \perp\!\!\!\perp_{\mathbb{Q}} V_{[m] \setminus \{j\}} | (H, X_j)$. Given an insert $\mathbb{Q}_{\alpha}^{X_j | W_j H}$, V_j is assumed to be distributed according to $\text{insert}(\mathbb{Q}_{\alpha}^{X_{n+1} | W_{n+1} H}, \mathbb{Q}^{W_{n+1} | H \square Y_{n+1} | X_{n+1}})$.

There seems to be one outstanding issue here: in general, we are interested in decision rules influencing X_j that may depend on the observations $V_{[n]}$ in addition to possible dependence on H and W_j , but we have only defined how consequences will be distributed given a subset of these inserts – namely, those that have no dependence on $V_{[n]}$.

However, by the mutual independence assumptions we make of consequences, we have for *any* insert α

$$\text{insert}(\mathbb{Q}_{\alpha}^{X_{n+1} | HW_{n+1} V_{[n]}}, \mathbb{Q}^{W_{n+1} V_{[n]} | H \square Y_{n+1} | X_{n+1}}) \quad (157)$$

$$= \mathbb{Q}^{W_{n+1} V_{[n]} | H_{n+1}} \odot \mathbb{Q}_{\alpha}^{X_{n+1} | HW_{n+1} V_{[n]}} \odot \mathbb{Q}^{Y_{n+1} | HW_{n+1} X_{n+1} V_{[n]}} \quad (158)$$

$$(159)$$

$$\mathbb{Q}^{W_j | X_j H}(w|a, h) = \mathbb{Q}^{W_1 | H}(w|h) \quad (160)$$

$$\implies \mathbb{Q}^{W_j | X_j H} = \mathbb{Q}^{W_1 | H} \otimes \text{erase}_X \quad (161)$$

Thus there exists $\mathbb{Q}^{W_j | H}$ for $j \in (n, m]$ and

$$\mathbb{Q}^{W_j | X_j} = \mathbb{Q}^{W_1 | H} \quad (162)$$

And by disintegration of $\mathbb{Q}^{W_j X_j Y_j | X_j H}$ (and Theorem 2.10)

$$\mathbb{Q}^{Y_j | X_j W_j H} = \mathbb{Q}^{Y_1 | X_1 W_1 H} \quad (163)$$

Thus the probability gap model \mathbb{Q} defined by the CBN features a collection of conditional probabilities:

- $\mathbb{Q}^{V_i|H}, i \in [n]$
- $\mathbb{Q}^{W_i|H}, i \in [m]$
- $\mathbb{Q}^{Y_i|X_i W_i H}, i \in [m]$

To this, we add the assumption of mutual conditional independence: $V_i \perp\!\!\!\perp_{\mathbb{Q}} V_{[m] \setminus \{i\}} | H$. Then we have

Representation of conditional IID models; can I just quote something here?

$$\mathbb{Q}^{V_{[n]} W_{(n,m)} | H} = \text{copy}_H^m \left(\left(\bigotimes_{i \in [n]} \mathbb{Q}^{V_i | H} \right) \otimes \left(\bigotimes_{i \in [m]} \mathbb{Q}^{W_i | H} \right) \right) \quad (164)$$

$$\mathbb{Q}^{Y_{(n,m)} | X_{(n,m)} W_{(n,m)} V_{[n]} | H} = \text{copy}_H^{m-n} \otimes \text{id}_{X W V} (\mathbb{Q}^{Y_i | X_i W_i H} \otimes \text{erase}_{V^n}) \quad (165)$$

$$\mathbb{Q}^{Y_i | H X_i V_{[n]} W_i} = \begin{array}{c} W_i \\ X_i \\ V_{[n]} \end{array} \begin{array}{c} \boxed{\mathbb{Q}^{Y_i | X_i W_i H}} \\ \hline * \end{array} \rightarrow Y_i \quad (166)$$

At this point, we note that $\mathbb{Q}^{V_{[n]} W_{(n,m)} | H}$ and $\mathbb{Q}^{Y_{(n,m)} | X_{(n,m)} W_{(n,m)} V_{[n]} | H}$ together define a probability 2-comb \mathbb{Q} , which motivates the following definition

Definition 4.1 (CBN probability 2-comb). A CBN probability 2-comb is a probability 2-comb $\mathbb{Q}^{V_{[n]} W_{(n,m)} | H \square Y_{(n,m)} | X_{(n,m)} | H}$ where $V_i = (W_i, X_i, Y_i)$ for $i \in [m]$, such that $V_i \perp\!\!\!\perp_{\mathbb{Q}}^2 V_{[m] \setminus \{i\}} | H$, for all $i \in [m]$

$$\mathbb{Q}^{V_i | H} = \mathbb{Q}^{V_j | H} \quad (167)$$

for $i, j \in [n]$,

$$\mathbb{Q}^{W_i | H} = \mathbb{Q}^{W_j | H} \quad (168)$$

for $i, j \in [m]$ and there exists some $\mathbb{Q}^{Y_j | X_j W_j H}, j \in [n]$ such that

$$\mathbb{Q}^{Y_i | X_i W_i H V_{<i}} = \mathbb{Q}^{Y_j | X_j W_j H} \otimes \text{erase}_{V^{i-1}} \quad (169)$$

For all $i \in (n, m]$, some $\mathbb{Q}^{Y_i | X_i W_i H V_{<i}}$.

Inserts for a CBN 2-comb take the form $\mathbb{Q}_\alpha^{X_{(n,m)} | V_{[n]} W_{(n,m)} | H}$. Such inserts are not necessarily decision rules as defined in the previous section – they don't necessarily admit an interpretation “if I see $(V_{[n]}, W_{(n,m)}, H)$ then I do this to $X_{(n,m)}$ ”, not the least because H is by definition unobserved. The question we will now ask is: under what conditions can we specify a see-do model for which the inserts *do* admit such an interpretation as decision rules and each decision rule corresponds to some insert \mathbb{Q}_α in our CBN.

4.2 See-do models compatible with causal Bayesian networks

When does a see-do model $\mathbb{T}^{V_{[n]}|H \square V_{(n,m)}|D}$ with decision rules $\{\mathbb{T}_\alpha^{D|V_{[n]}}\}_{\alpha \in A}$ correspond to a CBN probability 2-comb $\mathbb{Q}^{V_{[n]}W_{(n,m)}|H \square Y_{(n,m)}|X_{(n,m)}}$ with inserts $\{\mathbb{Q}_\alpha^{X_{(n,m)}|V_{[n]}W_{(n,m)}H}\}_{\alpha \in A}$?

By correspondence, we mean that for each $\alpha \in A$, the probabilistic model given by $\text{insert}(\mathbb{T}^{V_{[n]}|H \square V_{(n,m)}|D}, \mathbb{T}_\alpha^{D|V_{[n]}})$ followed by marginalising over D is the same as the model given by $\text{insert}(\mathbb{Q}^{V_{[n]}W_{(n,m)}|H \square Y_{(n,m)}|X_{(n,m)}}, \mathbb{Q}_\alpha^{X_{(n,m)}|V_{[n]}W_{(n,m)}H})$

$$\mathbb{T}_\alpha^{V_{[m]}|H} := \begin{array}{c} \text{Diagram showing } \mathbb{T}_\alpha^{V_{[m]}|H} \text{ as a causal model with nodes } H, V_{[n]}, D, V_{(n,m)}, X_{(n,m)}, Y_{(n,m)}, W_{(n,m)}. \end{array} \quad (170)$$

$$= \begin{array}{c} \text{Diagram showing } \mathbb{T}_\alpha^{V_{[m]}|H} \text{ as a causal model with nodes } H, V_{[n]}, X_{(n,m)}, Y_{(n,m)}, W_{(n,m)}. \end{array} \quad (171)$$

$$=: \mathbb{Q}_\alpha^{V_{[m]}|H} \quad (172)$$

Theorem 4.2. *Given a see-do model $\mathbb{T}^{V_{[n]}|H \square V_{(n,m)}|D}$ there exists a corresponding CBN probability 2-comb $\mathbb{Q}^{V_{[n]}W_{(n,m)}|H \square Y_{(n,m)}|X_{(n,m)}}$ if and only if*

1. $(V_{[n]}, W_j) \perp\!\!\!\perp_D^2 H$
2. $\mathbb{T}^{V_{[n]}W_j|H} = \mathbb{U}^{V_{[n]}W_j|H}$
3. $Y_j \perp\!\!\!\perp_D^2 D|W_jHX_j$
4. $\mathbb{T}^{Y_j|W_jHX_j} = \mathbb{T}^{Y_i|W_iHX_i}$ for $i \in [n], j \in (n, m)$

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

$$\mathbb{T}^{V_{[n]}V_j|HD} = \begin{array}{c} \text{Diagram showing } \mathbb{T}^{V_{[n]}V_j|HD} \text{ as a causal model with nodes } H, D, V_A, W_j, X_j, Y_j. \end{array} \quad (173)$$

For each $\mathbb{S}_\alpha^{D|V_{[n]}}$, define

$$\mathbb{R}_\alpha^{X_j|V_{[n]}W_jH} := \begin{array}{c} \text{Diagram showing } \mathbb{R}_\alpha^{X_j|V_{[n]}W_jH} \text{ as a causal model with nodes } W_j, H, V_A, X_j. \end{array} \quad (174)$$

Then

$$(175)$$

$$(176)$$

$$(177)$$

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

$$\mathbb{P}_d^{W_j|H}(w|h) = \mathbb{T}^{W_j|HD}(w|h, d) \quad (178)$$

$$\neq \mathbb{T}^{W_j|HD}(w|h, d') \quad (179)$$

$$= \mathbb{P}_{d'}^{W_j|H}(w|h) \quad (180)$$

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

$$\mathbb{P}_\alpha^{V[n]W_j|H} = \mathbb{T}^{V[n]W_j|H} \quad (181)$$

$$\neq \mathbb{U}^{V[n]W_j|H} \quad (182)$$

$$= \mathbb{Q}_\alpha^{V[n]W_j|H} \quad (183)$$

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

$$\mathbb{T}^{Y_j|W_j V[n] H X_j D}(y|w, v, h, x, d) \neq \mathbb{T}^{Y_j|W_j V[n] H X_j D}(y|w, v, h, x, d') \quad (184)$$

$$\text{and } \mathbb{T}^{X_j W_j V[n]|HD}(x, w, v|h, d) > 0 \quad (185)$$

$$\text{and } \mathbb{T}^{X_j W_j V[n]|HD}(x, w, v|h, d') > 0 \quad (186)$$

$$(187)$$

The latter conditions hold as if Equation 184 only held on sets of measure 0 then we could choose versions of the conditional probabilities such that the independence held.

Then

$$\mathbb{P}_d^{Y_j|W_jV_{[n]}HX_jD}(y|w, v, h, x) = \mathbb{T}^{Y_j|W_jV_{[n]}HX_jD}(y|w, v, h, x, d) \quad (188)$$

$$\neq \mathbb{T}^{Y_j|W_jV_{[n]}HX_jD}(y|w, v, h, x, d') \quad (189)$$

$$= \mathbb{P}_{d'}^{Y_j|W_jV_{[n]}HX_jD}(y|w, v, h, x) \quad (190)$$

$$\implies \mathbb{P}_d^{Y_j|W_jV_{[n]}HX_jD}(y|w, v, h, x) \neq \mathbb{Q}_d^{Y_j|W_jV_{[n]}HX_jD}(y|w, v, h, x) \quad (191)$$

$$\text{or } \mathbb{P}_{d'}^{Y_j|W_jV_{[n]}HX_jD}(y|w, v, h, x) \neq \mathbb{Q}_{d'}^{Y_j|W_jV_{[n]}HX_jD}(y|w, v, h, x) \quad (192)$$

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 $y \in Y$

$$\mathbb{P}_d^{Y_j|W_jV_{[n]}HX_jD}(y|w, v, h, x) = \mathbb{T}^{Y_j|W_jV_{[n]}HX_jD}(y|w, v, h, x) \quad (193)$$

$$\neq \mathbb{U}^{Y_j|W_jV_{[n]}HX_jD}(y|w, v, h, x) \quad (194)$$

$$\neq \text{model} Q_d^{Y_j|W_jV_{[n]}HX_jD}(y|w, v, h, x) \quad (195)$$

□

Conditional independences like $(V_{[n]}, W_j) \perp\!\!\!\perp_{\mathbb{T}} D|H$ and $Y_j \perp\!\!\!\perp_{\mathbb{T}} D|W_jV_{[n]}HX_j$ 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” $(W_j, V_{[n]}, H)$ and the “state” in Heckerman’s work crucial for defining limited unresponsiveness.

4.3 Proxy control

We say that $(V_{[n]}, W_j) \perp\!\!\!\perp_{\mathbb{T}} D|H$ expresses the notion that W_j is a *pre-choice variable* and $(W_j, V_{[n]}, X_j)$ 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 $Y_j \perp\!\!\!\perp H|W_jV_{[n]}X_j$ (i.e. the observed data allow us

to identify H for the purposes of determining $T^{Y_j|W_jV_{[n]}X_jH}$, then we can write

$$\begin{aligned}
T^{V_{[n]}V_{(n,m)}|HD} &= \begin{array}{c} \begin{array}{c} H \\ D \end{array} \begin{array}{c} \boxed{U^{W_jV_A|H}} \\ \boxed{T^{X|W_jV_AHD}} \end{array} \begin{array}{c} \xrightarrow{\quad} \\ \xrightarrow{\quad} \end{array} \begin{array}{c} \boxed{T^{Y_j|W_jV_AX_j}} \end{array} \begin{array}{c} V_A \\ W_j \\ Y_j \\ X_j \end{array} \end{array} \quad (196) \\
&= \begin{array}{c} \begin{array}{c} H \\ D \end{array} \begin{array}{c} \boxed{U^{W_jV_A|H}} \\ \boxed{T^{X|W_jV_AHD}} \end{array} \begin{array}{c} \xrightarrow{\quad} \\ \xrightarrow{\quad} \end{array} \boxed{K} \begin{array}{c} \xrightarrow{\quad} \\ \xrightarrow{\quad} \\ \xrightarrow{\quad} \\ \xrightarrow{\quad} \end{array} \begin{array}{c} V_A \\ W_j \\ Y_j \\ X_j \end{array} \end{array} = T^{V_{[n]}W_jX_j|HD}\mathbb{M} \quad (197)
\end{aligned}$$

That is, under conditions of full information, knowing how to control the proxies $(W_j, V_{[n]}, X_j)$ 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 $(Y_i, X_i, Z_i)_{i \in [n]}$ and a collection of potential outcomes $(Y_i^x)_{x \in X, i \in [n]}$. Also like causal Bayesian 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 $X : \Omega \rightarrow X$ and $\{Y^x : \Omega \rightarrow Y | x \in X\}$, define $Y^X : (\mathcal{Y}^x)_{x \in X}$. The selector $\pi : X \times Y^X \rightarrow Y$ is the function that sends $(x, y^1, \dots, y^{|X|}) \rightarrow y^x$.

Definition 5.2 (Potential outcomes: formal requirement). Given variables $Y : \Omega \rightarrow Y$ and $X : \Omega \rightarrow X$, we introduce a collection of latent variables called *potential outcomes* $Y^X := (\mathcal{Y}^x)_{x \in X}$ such that $Y = \pi \circ (X, Y^X)$. 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 $Y : \Omega \rightarrow Y$, $X : \Omega \rightarrow X$ and $W : \Omega \rightarrow W$, we can always define potential outcomes Y_X such that any consistent model $\mathbb{K}^{YX|W}$ can be extended to a consistent model of $\mathbb{K}^{YXY^X|W}$.*

Proof. Define $Y^X := (Y)_{x \in X}$. Then we can consistently extend $\mathbb{K}^{YX|W}$ to $\mathbb{K}^{YXY^X|W}$ by repeated application of Lemma 2.40. \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 $Y^x := \pi(x, Y)$ 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 $Y^x := \pi(x, Y)$ 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 $Y^{x'}$ 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 possible 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 $\mathbb{P}^{XY^X|H}$ of these variables as

The diagram shows a node H on the left. From H , three arrows point to three boxes: $\mathbb{P}^{Y^X|H}$, $\mathbb{P}^{Y|HY^X X}$, and $\mathbb{P}^{X|Y^X H}$. From $\mathbb{P}^{Y^X|H}$, an arrow points to Y^X . From $\mathbb{P}^{Y|HY^X X}$, an arrow points to Y . From $\mathbb{P}^{X|Y^X H}$, an arrow points to X . The equation (198) is shown to the right of the diagram.

For any collection of representative kernels $\mathbb{T}^{Y^X|H}$, $\mathbb{T}^{X|Y^X H}$ and $\mathbb{T}^{Y|HY^X X}$. We can simplify Equation 198 somewhat. Firstly, $\mathbb{P}^{Y|HY^X X}$ must always be represented a *selector kernel* $\Pi : X \times Y^{|X|} \rightarrow Y$, as shown by Lemma 5.5.

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

$$\mathbb{Q} := \begin{array}{c} Y^X \\ X \\ W \end{array} \begin{array}{c} \diagup \\ \diagdown \\ \longrightarrow \end{array} \Pi \longrightarrow Y \quad (199)$$

$$\iff \quad (200)$$

$$\mathbb{Q}_{(y^X, x, w)}^y = \Pi_{(x, y^X)}^y \quad \forall y, y^X, x, w \quad (201)$$

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

Proof. Recall $Y = \pi \circ (X, Y^X)$. Thus consistency implies that $Y \stackrel{a.s.}{=} \pi \circ (X, Y^X)$ for all $(x, y^X, w) \in \text{Range}(X) \times \text{Range}(Y) \times \text{Range}(W)$ such that $X^{-1}(x) \cap (Y^X)^{-1}(y^X) \cap W^{-1}(w) \neq \emptyset$. However, wherever $X^{-1}(x) \cap (Y^X)^{-1}(y^X) \cap W^{-1}(w) = \emptyset$, consistency implies $\mathbb{T}^{YY^X X|W}(y, y^X, x|w) = 0$ and so $\mathbb{T}^{Y|Y^X X W}$ is arbitrary on this collection of values. Equations 199 and 201 are equivalent to the statement $Y \stackrel{a.s.}{=} \pi \circ (X, Y^X)$. \square

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

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

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

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

$$\mathbb{T}^{Y|HWXD}(y|h, w, 0, 0) \neq \mathbb{T}^{Y|HWXD}(y|h, w, 0, 2) \quad (202)$$

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

$$\mathbb{P}_0^{Y|HD}(y|h, 0) = \sum_{x \in \{0, 1\}, w \in W} \mathbb{T}^{W|H}(w|h) \mathbb{S}_0^{D|W}(0|w) \mathbb{T}^{X|HWD}(x|h, w, 0) \mathbb{T}^{Y|HWXD}(y|h, w, x, 0) \quad (203)$$

$$= \mathbb{T}^{Y|HWXD}(y|h, w, 0, 0) \quad (204)$$

$$\neq \mathbb{P}_2^{Y|HD} \quad (205)$$

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

assumption

$$\mathbb{P}_0^{\mathbf{Y}|\mathbf{H}\mathbf{D}}(y|h, 0) = \sum_{y^X \in Y^2, x \in \{0,1\}} \mathbb{T}^{\mathbf{Y}^X|\mathbf{H}}(y^X|h) \mathbb{T}^{\mathbf{X}|\mathbf{H}\mathbf{D}\mathbf{Y}^X}(x|h, 0, y^X) \Pi(y|x, y^X) \quad (206)$$

$$= \sum_{y^X} \mathbb{T}^{\mathbf{Y}^X|\mathbf{H}}(y^X|h) \Pi(y|0, y^X) \quad (207)$$

$$= \sum_{y^X \in Y^2, x \in \{0,1\}} \mathbb{T}^{\mathbf{Y}^X|\mathbf{H}}(y^X|h) \mathbb{T}^{\mathbf{X}|\mathbf{H}\mathbf{D}\mathbf{Y}^X}(x|h, 2, y^X) \Pi(y|x, y^X) \quad (208)$$

$$= \mathbb{P}_2^{\mathbf{Y}|\mathbf{H}\mathbf{D}} \quad (209)$$

Here we use the property $\mathbf{Y}^X \perp\!\!\!\perp_{\mathbb{T}} \mathbf{D}|\mathbf{H}$, implied by the assumption that \mathbf{Y}^X is a prechoice variable. Equations 205 and 209 are clearly contradictory, thus there can be no potential outcomes \mathbf{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 additional 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 $\{\mathbb{T}^{\mathbf{WZ}|\mathbf{H}\mathbf{D}}, \{\mathbb{S}_\alpha\}_{\alpha \in A}\}$, and let \mathbf{U} be the sequence of all prechoice variables. For some \mathbf{Y} and \mathbf{X} , there exist potential outcomes \mathbf{Y}^X in the sense of Definition 5.6 if and only if $\mathbb{T}^{\mathbf{Y}|\mathbf{U}\mathbf{X}}$ exists and is deterministic.*

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

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

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

$$\mathbf{Y} \perp\!\!\!\perp_{\mathbb{T}} \mathbf{D}|\mathbf{W}\mathbf{X} \quad (210)$$

Proof. $\mathbb{T}^{\mathbf{Y}|\mathbf{U}\mathbf{X}}$ exists only if $\mathbf{Y} \perp\!\!\!\perp_{\mathbb{T}} \mathbf{D}|\mathbf{U}\mathbf{X}$. \square

Note the similarity between Equation 210 and the condition for proxy control in the previous section. Indeed, the two are identical if we identify \mathbf{U} with $(\mathbf{W}_j, \mathbf{V}_A, \mathbf{X}_j)$.

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 $\mathbb{S} : X \rightarrow \Delta(D)$. We have a utility function $u : Y \rightarrow \mathbb{R}$ 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 \mathbf{X}, \mathbf{Y} and \mathbf{D} such that $\mathbf{X}_{x\mathbf{d}y} = x$, $\mathbf{Y}_{x\mathbf{d}y} = y$ and $\mathbf{D}_{x\mathbf{d}y} = d$. Then making the following additional assumptions:

1. *Holding the hypothesis h fixed the observations as have the same distribution under any strategy: $\mathbb{P}_{h,\mathbb{S}}[\mathbf{X}] = \mathbb{P}_{h,\mathbb{S}'}[\mathbf{X}]$ 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: $\mathbb{S} = \mathbb{P}_{h,\mathbb{S}}[\mathbf{D}|\mathbf{X}]$*
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 $\mathbb{P}_{h,\mathbb{Q}}[\mathbf{Y}|\mathbf{D}\mathbf{X}] = \mathbb{P}_{h,\mathbb{S}}[\mathbf{Y}|\mathbf{D}\mathbf{X}]$ (our strategy tells us nothing about the consequences that we don’t already know from the observations and decisions)*

Then there exists a unique see-do model $(\mathbb{T}, \mathbf{H}', \mathbf{D}', \mathbf{X}', \mathbf{Y}')$ such that $\mathbb{P}_{h,\mathbb{S}}[\mathbf{X}\mathbf{D}\mathbf{Y}]^{ijk} = \mathbb{T}[\mathbf{X}'|\mathbf{H}']_h^i \mathbb{S}_i^j \mathbb{T}[\mathbf{Y}'|\mathbf{X}'\mathbf{H}'\mathbf{D}']_{ijk}^k$.

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

$$\mathbb{P}[\mathbf{X}\mathbf{D}\mathbf{Y}]^{ijk} = \mathbb{P}[\mathbf{X}]^i \mathbb{P}[\mathbf{D}|\mathbf{X}]_i^j \mathbb{P}[\mathbf{Y}|\mathbf{X}\mathbf{D}]_{ij}^k \quad (211)$$

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

$$\mathbb{T}_{hj}^{kl} = \mathbb{P}_{h,\mathbb{S}}[\mathbf{X}]^k \mathbb{P}_{h,\mathbb{S}}[\mathbf{Y}|\mathbf{X}\mathbf{D}]_{kj}^l \quad (212)$$

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

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

We then have

$$\mathbb{T}[\mathbf{X}'|\mathbf{H}'\mathbf{D}']_{hj}^k = \mathbb{T}\mathbf{X}'_{hj}^k \quad (213)$$

$$= \sum_l \mathbb{T}_{hj}^{kl} \quad (214)$$

$$= \mathbb{P}_{h,\mathbb{S}}[\mathbf{X}]^k \quad (215)$$

$$= \mathbb{T}[\mathbf{X}'|\mathbf{H}'\mathbf{D}']_{hj'}^k \quad (216)$$

Thus $\mathbf{X}' \perp\!\!\!\perp_{\mathbb{T}} \mathbf{D}'|\mathbf{H}'$ and so $\mathbb{T}[\mathbf{X}'|\mathbf{H}']$ exists (section 2.10.2) and $(\mathbb{T}, \mathbf{H}', \mathbf{D}', \mathbf{X}', \mathbf{Y}')$ is a see-do model.

Applying Equation 211 to $\mathbb{P}_{h,\mathbb{S}}$:

$$\mathbb{P}_{h,\mathbb{S}}[\mathbf{XDY}]^{ijk} = \mathbb{P}_{h,\mathbb{S}}[\mathbf{X}]^i \mathbb{P}_{h,\mathbb{S}}[\mathbf{D}|\mathbf{X}]_i^j \mathbb{P}_{h,\mathbb{S}}[\mathbf{Y}|\mathbf{XD}]_{ij}^k \quad (217)$$

$$= \mathbb{P}_{h,\mathbb{S}}[\mathbf{X}]^i \mathbb{P}_{h,\mathbb{S}}[\mathbf{Y}|\mathbf{XD}]_{ij}^k \quad (218)$$

$$= \mathbb{P}_{h,\mathbb{S}}[\mathbf{D}|\mathbf{X}]_i^j \mathbb{T}[\mathbf{X}'\mathbf{Y}'|\mathbf{H}'\mathbf{D}']_{hj}^{ik} \quad (219)$$

$$= \mathbb{S}_i^j \mathbb{T}[\mathbf{X}'\mathbf{Y}'|\mathbf{H}'\mathbf{D}']_{hj}^{ik} \quad (220)$$

$$= \mathbb{S}_i^j \mathbb{T}[\mathbf{X}'|\mathbf{H}'\mathbf{D}']_{hj}^i \mathbb{T}[\mathbf{Y}'|\mathbf{X}'\mathbf{H}'\mathbf{D}']_{ihj}^k \quad (221)$$

$$= \mathbb{T}[\mathbf{X}'|\mathbf{H}']_h^i \mathbb{S}_i^j \mathbb{T}[\mathbf{Y}'|\mathbf{X}'\mathbf{H}'\mathbf{D}']_{ihj}^k \quad (222)$$

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

$$\mathbb{P}_{h,\mathbb{S}}[\mathbf{X}]^i = \mathbb{P}_{h,\mathbb{Q}}[\mathbf{X}]^i \quad (223)$$

$$\mathbb{P}_{h,\mathbb{S}}[\mathbf{Y}|\mathbf{XD}]_{ij}^k = \mathbb{P}_{h,\mathbb{Q}}[\mathbf{Y}|\mathbf{XD}]_{ij}^k \text{ for some version of } \mathbb{P}_{h,\mathbb{Q}}[\mathbf{Y}|\mathbf{XD}] \quad (224)$$

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

$$\mathbb{T}_{hj}^{kl} = \mathbb{P}_{h,\mathbb{Q}}[\mathbf{X}]^k \mathbb{P}_{h,\mathbb{Q}}[\mathbf{Y}|\mathbf{XD}]_{kj}^l \quad (225)$$

Then by substitution of \mathbb{Q} for \mathbb{S} in Equation 217 and working through the same steps

$$\mathbb{P}_{h,\mathbb{S}}[\mathbf{XDY}]^{ijk} = \mathbb{T}[\mathbf{X}'|\mathbf{H}']_h^i \mathbb{Q}_i^j \mathbb{T}[\mathbf{Y}'|\mathbf{X}'\mathbf{H}'\mathbf{D}']_{ihj}^k \quad (226)$$

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

7 Appendix: Counterfactual representation

Definition 7.1 (Parallel potential outcomes). Given a Markov kernel space (\mathbb{K}, E, F) , a collection of variables $\{\mathbf{Y}_i, \mathbf{Y}(W), \mathbf{W}_i\}$, $i \in [n]$, where \mathbf{Y}_i and $\mathbf{Y}(W)$ are random variables and \mathbf{W}_i could be either a state or random variables is a *parallel potential outcome submodel* if $\mathbb{K}[\mathbf{Y}_i|\mathbf{W}_i\mathbf{Y}(W)]$ exists and $\mathbb{K}[\mathbf{Y}_i|\mathbf{W}_i\mathbf{Y}(W)]_{kj_1j_2\dots j_{|\mathbf{W}|}} = \delta[j_k]$.

How this will change: a parallel potential outcomes model is a comb
 $\mathbb{E}[Y(W)|H] \Rightarrow \mathbb{E}[Y_i|W_i Y(W)]$.

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 $W_i = w_i$ then the corresponding outcome $Y_i \stackrel{a.s.}{=} Y(w_i)$. 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 $Y_i \stackrel{a.s.}{=} Y(w_i)$ 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 $Y(w)$, 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 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

Exchangeable 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 $\mathbb{P}[X_1 X_2] = \mathbb{P}[X_2 X_1]$. 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 $\mathbb{P}[Y_1] = \mathbb{P}[Y_2]$ 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 $\mathbb{P}[Y_2|Y_1] = \mathbb{P}[Y_1|Y_2]$, then our model of subjective uncertainty is exchangeable.

De Finetti’s representation theorem establishes the fact that any infinite exchangeable sequence Y_1, Y_2, \dots can be modeled by the product of a *prior* probability $\mathbb{P}[J]$ with J taking values in the set of marginal probabilities $\Delta(Y)$ and a conditionally independent and identically distributed Markov kernel $\mathbb{P}[Y_A|J]_j^{y_A} = \prod_{i \in A} \mathbb{P}[Y_i|J]_j^{y_i}$.

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 analogous 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 exchangeability 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 $((D_i, Y_i))_{i \in [n]}$ with Y_i random variables is *exchangeable* conditional on Z if, defining $Y_{[n]} = (Y_i)_{i \in [n]}$ and $D_{[n]} = (D_i)_{i \in [n]}$, $\mathbb{K}[Y_{[n]}|D_{[n]}Z]$ exists and for any bijection $\pi : [n] \rightarrow [n]$ $\mathbb{K}[Y_{\pi([n])}|D_{\pi([n])}Z] = \mathbb{K}[Y_{[n]}|D_{[n]}Z]$.

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 $\mathbb{K}'[X|U]$ exists and $\mathbb{K}'[X|U] = \mathbb{K}$.

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 $Y \circ X$ on (\mathbb{K}', E', F') and any state variable D with $D \circ U$ on (\mathbb{K}', E', F') and under this identification $\mathbb{K}'[Y \circ X|D \circ U]$ exists iff $\mathbb{K}[Y|D]$ exists and $\mathbb{K}'[Y \circ X|D \circ U] = \mathbb{K}[Y|D]$. 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 $Y := Y \circ X$ and $D := D \circ U$.

I think this is a very standard thing to do – propose some X and $\mathbb{P}(X)$ then introduce some random variable Y and $\mathbb{P}(XY)$ 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 $((D_i, Y_i))_{i \in [n]}$ 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 $((D_i, Y_i))_{i \in \mathbb{N}}$ is exchangeable conditional on Z .

Here that we identify Z and $((D_i, Y_i))_{i \in [n]}$ defined on the extension with the original variables defined on (\mathbb{K}, E, F) while $((D_i, Y_i))_{i \in \mathbb{N} \setminus [n]}$ 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 concerning 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 $((D_i, Y_i))_{i \in [n]}$ with Y_i random variables is *deterministically reproducible* conditional on Z if $n \geq 2$, $\mathbb{K}[Y_{[n]}|D_{[n]}Z]$ exists and $\mathbb{K}[Y_{\{i,j\}}|D_{\{i,j\}}Z]_{kk}^{lm} = \llbracket l = m \rrbracket \mathbb{K}[Y_i|D_iZ]_k^l$ 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 $((D_i, Y_i))_{i \in [n]}$ with $n \geq 2$ and a conditioning variable Z , (\mathbb{K}, E, F) can be extended with a set of variables $Y(D) := (Y(i))_{i \in D}$ such that $\{Y_i, Y(D), D_i\}$ is a parallel potential outcome submodel if and only if $((D_i, Y_i))_{i \in [n]}$ is extendably exchangeable and deterministically reproducible conditional on Z .*

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

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

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

$$\mathbb{K}[X_n|D_nZ]_{dz}^x = \sum_{y \in Y^{|D|}} \mathbb{K}[X_{[n]}|D_nZY(D)]_{dzy}^x \mathbb{K}[Y(D)|D_{[n]}Z]_{dz}^y \quad (227)$$

$$= \sum_{y \in Y^{|D|}} \prod_{i \in [n]} \delta[y_{d_i}]^{x_i} \mathbb{K}[Y(D)|D_nZ]_{dz}^y \quad (228)$$

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 $\mathbb{K}[Y_{[n]}|D_{[n]}Z]_{dz}^x = 0$ 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 $Q_d \subset [n] := \{i \mid \nexists i \in [n] : j < i \text{ \& } d_j = d_i\}$, 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 $R_d \subset D : \{d_i \mid i \in Q_d\}$ 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 $y' = (y_i)_{i \in Q_d^C}$, $x_{Q_d} = (x_i)_{i \in Q_d}$, $Y(R_d) = (Y_d)_{d \in R_d}$ and $Y(S_d) = (Y_d)_{d \in S_d}$.

$$\mathbb{K}[X_{[n]}|D_{[n]}Z]_{dz}^x = \sum_{y \in Y^{|D|}} \prod_{i \in Q_d} \delta[y_{d_i}]^{x_i} \mathbb{K}[Y(D)|D_{[n]}Z]_{dz}^y \quad (229)$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbb{K}[Y(R_d)Y(R_d^C)|D_{Q_d}D_{Q_d^C}Z]_{d_{Q_d}d_{Q_d^C}z}^{x_{Q_d}y'} \quad (230)$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbb{K}[Y_{R_d}Y_{R_d^C}|D_{Q_d}D_{Q_d^C}Z]_{dz}^{x_{Q_d}y'} \quad (231)$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbb{K}[Y_{[n]}|D_{[n]}Z]_{dz}^{x_{Q_d}y'} \quad (\text{using exchangeability}) \quad (232)$$

Note that

Only if: We aim to show that the sequences $Y_{[n]}$ and $D_{[n]}$ in a parallel potential outcomes submodel are exchangeable and deterministically reproducible. \square

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

$$I_{K..} := I_K \setminus I_L \setminus I_J \quad (233)$$

$$I_{KL.} := I_K \cap I_L \setminus I_J \quad (234)$$

$$I_{K..J} := I_K \cap I_J \setminus I_L \quad (235)$$

$$I_{KLJ} := I_K \cap I_L \cap I_J \quad (236)$$

$$I_{..L} := I_L \setminus I_K \setminus I_J \quad (237)$$

$$I_{..LJ} := I_L \cap I_J \setminus I_K \quad (238)$$

$$I_{..J} := I_J \setminus I_K \setminus I_L \quad (239)$$

$$O_{K..} := O_K \setminus I_N \setminus I_J \quad (240)$$

$$O_{KL.} := O_K \cap I_L \setminus I_J \quad (241)$$

$$O_{K..J} := O_K \cap I_J \setminus I_L \quad (242)$$

$$O_{KLJ} := O_K \cap I_L \cap I_J \quad (243)$$

$$O_{..L} := O_L \setminus I_J \quad (244)$$

$$O_{..LJ} := O_L \cap I_J \quad (245)$$

Also define

$$(\mathbb{P}, I_P, O_P) := \mathbb{K} \rightrightarrows \mathbb{L} \quad (246)$$

$$(\mathbb{Q}, I_Q, O_Q) := \mathbb{L} \rightrightarrows \mathbb{J} \quad (247)$$

Then

$$(\mathbb{K} \Rightarrow \mathbb{L}) \Rightarrow \mathbb{J} = \mathbb{P} \Rightarrow \mathbb{J} \quad (248)$$

$$= \begin{array}{c} \text{Diagram with boxes } \mathbb{P} \text{ and } \mathbb{J} \\ \text{Inputs: } I_{P\cdot}, I_{PJ}, I_{\cdot J} \\ \text{Outputs: } O_{P\cdot}, O_{PJ}, O_J \end{array} \quad (249)$$

$$= \begin{array}{c} \text{Diagram with boxes } \mathbb{K}, \mathbb{L}, \text{ and } \mathbb{J} \\ \text{Inputs: } I_{K\cdot}, I_{KL\cdot}, I_{\cdot L}, I_{K\cdot J}, I_{KLJ}, I_{\cdot LJ}, I_{\cdot\cdot J} \\ \text{Outputs: } O_{K\cdot}, O_{KL\cdot}, O_{K\cdot J}, O_{KLJ}, O_{\cdot L}, O_{\cdot LJ}, O_J \end{array} \quad (250)$$

$$\stackrel{\text{perm}}{=} \begin{array}{c} \text{Diagram with boxes } \mathbb{K}, \mathbb{L}, \text{ and } \mathbb{J} \\ \text{Inputs: } I_{K\cdot}, I_{KL\cdot}, I_{K\cdot J}, I_{KLJ}, I_{\cdot L}, I_{\cdot LJ}, I_{\cdot\cdot J} \\ \text{Outputs: } O_{K\cdot}, O_{KL\cdot}, O_{K\cdot J}, O_{KLJ}, O_{\cdot L}, O_{\cdot LJ}, O_J \end{array} \quad (251)$$

$$= \begin{array}{c} \text{Diagram with boxes } \mathbb{K} \text{ and } \mathbb{Q} \\ \text{Inputs: } I_{K\cdot}, I_{KQ}, I_{\cdot Q} \\ \text{Outputs: } O_{K\cdot}, O_{KQ}, O_Q \end{array} \quad (252)$$

$$= \mathbb{K} \Rightarrow (\mathbb{L} \Rightarrow \mathbb{J}) \quad (253)$$

9 Appendix: String Diagram Examples

Recall the definition of *connection*:

Definition 9.1 (Connection).

$$\mathbb{K} \Rightarrow \mathbb{L} := \begin{array}{c} \text{Diagram with boxes } \mathbb{K} \text{ and } \mathbb{L} \\ \text{Inputs: } I_{F\cdot}, I_{FS}, I_{\cdot S} \\ \text{Outputs: } O_{F\cdot}, O_{FS}, O_S \end{array} \quad (254)$$

$$:= \mathbb{J} \quad (255)$$

$$\mathbb{J}_{yqr}^{zxw} = \mathbb{K}_{yq}^{zx} \mathbb{L}_{xqr}^w \quad (256)$$

Equation 254 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:

$$(\mathbb{J}, (\mathbb{I}_{F\cdot}, \mathbb{I}_{FS}, \mathbb{I}_S), (\mathbb{O}_{F\cdot}, \mathbb{O}_{FS}, \mathbb{O}_S)) = \left[\begin{array}{c} \mathbb{I}_{F\cdot} \\ \mathbb{I}_{FS} \\ \mathbb{I}_S \end{array} \right] \left[\begin{array}{c} \mathbb{K} \\ \mathbb{L} \end{array} \right] \left[\begin{array}{c} \mathbb{O}_{FS} \\ \mathbb{O}_{F\cdot} \end{array} \right] \quad (257)$$

$$(258)$$

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 $\mathbb{K} : X \times \mathcal{Y} \rightarrow [0, 1]$ 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* $\delta : X \rightarrow \Delta(\mathcal{X})$ is the Markov kernel defined by

$$\delta(x, A) = \begin{cases} 1 & x \in A \\ 0 & \text{otherwise} \end{cases} \quad (259)$$

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

$$\text{del}_X := \text{---} * \text{copy}_X \quad := \text{---} \bullet \text{---} \quad (260)$$

and satisfying the commutative comonoid equations

$$\text{---} \bullet \text{---} \bullet \text{---} = \text{---} \bullet \text{---} \bullet \text{---} \quad (261)$$

$$\text{---} \bullet \text{---} * = \text{---} = \text{---} \bullet \text{---} * \quad (262)$$

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \\ \text{---} \bullet \begin{array}{l} \searrow \\ \nearrow \end{array} \end{array} = \begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \\ \text{---} \bullet \begin{array}{l} \searrow \\ \nearrow \end{array} \end{array} \quad (263)$$

as well as compatibility with the monoidal structure

$$\begin{array}{c} X \otimes Y \text{---} * \\ = X \text{---} * \end{array} \quad (264)$$

$$\begin{array}{c} X \otimes Y \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \\ \text{---} \bullet \begin{array}{l} \searrow \\ \nearrow \end{array} \end{array} \begin{array}{l} X \otimes Y \\ X \otimes Y \end{array} = \begin{array}{c} X \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \\ \text{---} \bullet \begin{array}{l} \searrow \\ \nearrow \end{array} \end{array} \begin{array}{l} X \\ Y \\ X \\ Y \end{array} \quad (265)$$

and the naturality of del , which means that

$$\begin{array}{c} \text{---} \boxed{f} \text{---} * \\ = \text{---} * \end{array} \quad (266)$$

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

$$\mathbb{K}\mathbb{L} : \mathsf{X} \rightarrow \Delta(\mathsf{Z}) \quad (267)$$

$$\text{defined by } (\mathbb{K}\mathbb{L})(x, A) = \int_{\mathsf{Y}} \mathbb{L}(y, A) \mathbb{K}(x, dy) \quad (268)$$

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

$$\mathbb{K} \otimes \mathbb{L} : (\mathsf{X}, \mathsf{W}) \rightarrow \Delta(\mathsf{Y}, \mathsf{Z}) \quad (269)$$

$$\text{defined by } \mathbb{K} \otimes \mathbb{L}(x, w, A \times B) = \mathbb{K}(x, A) \mathbb{L}(w, B) \quad (270)$$

The identity map is

$$\text{Id}_{\mathsf{X}} : \mathsf{X} \rightarrow \Delta(\mathsf{X}) \quad (271)$$

$$\text{defined by } (\text{Id}_{\mathsf{X}})(x, A) = \delta(x, A) \quad (272)$$

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

The swap map is given by

$$\text{swap}_{\mathbf{X}, \mathbf{Y}} : (\mathbf{X}, \mathbf{Y}) \rightarrow \Delta(\mathbf{Y}, \mathbf{X}) \quad (273)$$

$$\text{defined by } (\text{swap}_{\mathbf{X}, \mathbf{Y}})(x, y, A \times B) = \delta(x, B)\delta(y, A) \quad (274)$$

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 $(\mathbf{X}, (\mathbf{Y}, \mathbf{Z})) \cong ((\mathbf{X}, \mathbf{Y}), \mathbf{Z})$.

The copy map is given by

$$\text{copy}_{\mathbf{X}} : \mathbf{X} \rightarrow \Delta(\mathbf{X}, \mathbf{X}) \quad (275)$$

$$\text{defined by } (\text{copy}_{\mathbf{X}})(x, A \times B) = \delta_x(A)\delta_x(B) \quad (276)$$

and the erase map by

$$\text{del}_{\mathbf{X}} : \mathbf{X} \rightarrow \Delta(*) \quad (277)$$

$$\text{defined by } (\text{del}_{\mathbf{X}})(x, A) = \delta(*, A) \quad (278)$$

$$(279)$$

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 261 to 266 are known to be satisfied for the underlying unlabeled Markov kernels. We need to show is that they hold given our stricter criterion of labeled Markov kernel equality; that the underlying kernels *and the label sets* match. It is sufficient to check the label sets only.

□

References

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. Interditions, 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 Bojaczek 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>.

- Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 1 edition edition, May 2018. ISBN 978-0-465-09760-9.
- 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 Shohar. 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: