

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

David Johnston

December 21, 2021

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Variables and Probability Models</b>	<b>3</b>
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 sets . . . . .	14
2.7	Probability sets defined by marginal and conditional probabilities	16
2.8	Probability gap models . . . . .	20
2.8.1	Disintegrations . . . . .	22
2.8.2	Conditional independence . . . . .	24
2.8.3	Extended conditional independence . . . . .	25
2.8.4	Graphical properties of conditional independence . . . . .	28
2.9	Results I use that don't really fit into the flow of the text . . . .	28
2.9.1	Repeated variables . . . . .	28
<b>3</b>	<b>Decision theoretic causal inference</b>	<b>30</b>
3.1	Decision problems . . . . .	30
3.2	Decisions as measurement procedures . . . . .	32
3.3	Causal models similar to see-do models . . . . .	33
3.4	See-do models and classical statistics . . . . .	34
<b>4</b>	<b>Repeatable experiments</b>	<b>35</b>
4.1	Assumptions of repeatability applicable to models of decisions and consequences . . . . .	37
4.2	Representations of contractible probability models . . . . .	40

4.3	Extending contractible do models with observations . . . . .	46
<b>5</b>	<b>Causal Bayesian Networks</b>	<b>49</b>
5.1	Probability 2-combs represented by causal Bayesian networks . .	50
5.2	See-do models corresponding to causal Bayesian networks . . . .	52
5.3	Proxy control . . . . .	55
<b>6</b>	<b>Potential outcomes</b>	<b>56</b>
<b>7</b>	<b>Appendix: see-do model representation</b>	<b>61</b>
<b>8</b>	<b>Appendix: Counterfactual representation</b>	<b>62</b>
8.1	Parallel potential outcomes representation theorem . . . . .	63
<b>9</b>	<b>Appendix: Connection is associative</b>	<b>66</b>
<b>10</b>	<b>Appendix: String Diagram Examples</b>	<b>67</b>
<b>11</b>	<b>Markov variable maps and variables form a Markov category</b>	<b>68</b>

## 1 Introduction

Two widely used approaches to causal modelling are *graphical causal models* and *potential outcomes models*. Graphical causal models, which include Causal Bayesian Networks and Structural Causal Models, provide a set of *intervention* operations that take probability distributions and a graph and return a modified probability distribution (Pearl, 2009). Potential outcomes models feature *potential outcome variables* that represent the “potential” value that a quantity of interest would take under the right circumstances, a potential that may be realised if the circumstances actually arise, but will otherwise remain only a potential or *counterfactual* value (Rubin, 2005).

One challenge for both of these approaches is understanding how their causal primitives – interventions and potential outcome variables respectively – relate to the causal questions we are interested in. This challenge is related to the distinction, first drawn by (Korzybski, 1933), between “the map” and “the territory”. Causal models, like other models, are “maps” that purport to represent a “territory” that we are interested in understanding. Causal primitives are elements of the maps, and the things to which they refer are parts of the territory. The maps contain all the things that we can talk about unambiguously, so it is challenging to speak clearly about how parts of the maps relate to parts of the territory that fall outside of the maps.

For example, Hernán and Taubman (2008), who observed that many epidemiological papers have been published estimating the “causal effect” of body mass index and argued that, because *actions* affecting body mass index<sup>1</sup> are vaguely

---

<sup>1</sup>the authors use the term “intervention”, but they do not use it mean a formal operation on a graphical causal model, and we reserve the term for such operations to reduce ambiguity.

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  $\Omega$ . 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  $\Omega$ , 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 deci-

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

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

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

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

**Definition 2.2** (Markov kernel (discrete sets)). A Markov kernel  $\mathbb{K} : X \rightarrow Y$  is a matrix  $(\mathbb{K}(y|x))_{x \in X, y \in Y} \in [0, 1]^{|X| \times |Y|}$  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:

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

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:

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

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:

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

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

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

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

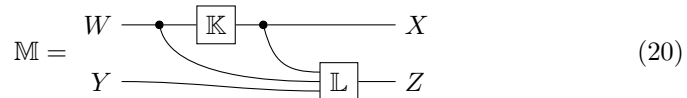
$$= \mathbb{K}^n(z|(y)^n) \quad (19)$$

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

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

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

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

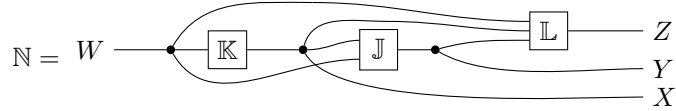
$$(\text{id}_X \otimes \mathbb{L})](A \times B|w, w, y) \quad (24)$$

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

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

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

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  $\mathbb{M}$ :



$$\mathbb{N} = W \text{ --- } \text{[Diagram]} \quad (28)$$

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. No actual procedure can be guaranteed to return elements of a mathematical set known in advance – anything can fail – but we assume that we can study procedures reliable enough that we don’t lose much by making this assumption.

**Definition 2.3** (Measurement procedure). A *measurement procedure* is a procedure that involves interacting with the real world somehow and delivering an element of a mathematical set as a result. The set of possible values is known prior to the measurement taking place, but the value that it will yield is not known. A procedure is given the font  $\mathcal{B}$ , we say it takes values in  $X$  and  $\mathcal{B} \bowtie x$  is the proposition that the procedure  $\mathcal{B}$  will yield the value  $x \in X$ .  $\mathcal{B} \bowtie A$  for  $A \subset X$  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  $x \in B$  (note that  $\mathcal{B}$  and  $\mathcal{C}$  could involve different actions in the real world).

Measurement procedures are like functions without well-defined domains. We can compose measurement procedures with functions to produce new measurement procedures.

**Definition 2.4** (Composition of functions with procedures). Given 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 any procedure  $\mathcal{C}$  that yields  $f(x)$  whenever  $\mathcal{B}$  yields  $x$ . We can construct such a procedure by describing the steps: first, do  $\mathcal{B}$  and secondly, apply  $f$  to the value yielded by  $\mathcal{B}$ .

For example,  $\mathcal{MA}$  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. Measurement procedure composition is associative:

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

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

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

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

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. Unlike composition, we can’t easily reason about such an operation mathematically, because of the fact that measurement procedures have a foot in the real world. Our approach here is to suppose that there is some master measurement procedure  $\mathcal{S}$  which takes values in  $\Psi$  that handles all of the “real world” interaction relevant to our problem. Specifically, we assume that any measurement procedure of interest to our problem can be written as the composition  $f \circ \mathcal{S}$  for some  $f$ .

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

Thus far we have defined by  $\Psi$  a “fundamental probability set” limited to observable variables – which is to say, limited to variables that are associated with measurement procedures. Unobserved variables need not be associated with measurement procedures, and to accommodate these we use instead of  $\Psi$  a richer fundamental probability set  $\Omega$  which represents both observed and unobserved variables.

**Definition 2.5** (Fundamental probability set). The fundamental probability set (or *sample space*)  $\Omega$  is a measurable set with  $\sigma$ -algebra  $\mathcal{F}$ .

Observables are represented by a function  $S : \Omega \rightarrow \Psi$ , and values of  $\omega$  are related to propositions about measurement procedures via the criterion of *consistency with observation*.

**Definition 2.6** (Consistency with observation). An element  $\omega \in \Omega$  is *consistent with observation* if the result yielded by  $S \bowtie S(\omega)$

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

different sets of measurement outcomes  $\Psi$  and  $\Psi'$  entail a different measurement procedures  $\mathcal{S}$  and  $\mathcal{S}'$ , but different fundamental probability sets  $\Omega$  and  $\Omega'$  may be used to model a single procedure  $\mathcal{S}$ .

As far as we know, distinguishing variables from procedures is somewhat non-standard, but we feel it is useful to distinguish the formal elements of our theory (variables) from the semi-formal elements (measurement procedures). 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  $f \circ \mathcal{S}$  for some  $f$  – is modeled by a variable, but there may be variables defined on  $\Omega$  that do not factorise through  $\mathcal{S}$ , and these variables do not model procedures.

## 2.4 Events

To recap, we have a procedure  $\mathcal{S}$  yielding values in  $\Psi$  that measures everything we are interested in, a fundamental probability set  $\Omega$  and a function  $\mathcal{S}$  that models  $\mathcal{S}$  in the sense of Definition 2.6. We assume also that  $\Psi$  has a  $\sigma$ -algebra  $\mathcal{E}$  (this may be the power set of  $\Psi$ , as measurement procedures are typically limited to finite precision).  $\Omega$  is equipped with a  $\sigma$ -algebra  $\mathcal{F}$  such that  $\sigma(\mathcal{S}) \subset \mathcal{F}$ . If a procedure  $\mathcal{X} = f \circ \mathcal{S}$  then we define  $X : \Omega \rightarrow X$  by  $X := f \circ \mathcal{S}$ .

If a particular procedure  $\mathcal{X} = f \circ \mathcal{S}$  eventually yields a value  $x$ , then the values of  $\Omega$  consistent with observation must be a subset of  $X^{-1}(x)$ . We define an *event*  $X \bowtie x \equiv X^{-1}(x)$ , which we read “the event that  $X$  yields  $x$ ”. An event  $X \bowtie x$  occurs if the consistent values of  $\Omega$  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  $(\Omega, \mathcal{F})$  along with our collection of variables is a “model skeleton” – it tells us what kind of data we might see. The process  $S$  which tells us which part of the world we’re interested in is related to the model  $\Omega$  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  $\mu$  on  $\Omega$  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, I need a model that 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 impose some additional conditions on probability gap models. In particular, a probability gap model imposes some requirements on the final probability distributions of interest; that is, it defines a *probability set*, which is a subset of probability models on  $(\Omega, \mathcal{F})$ . Probability gaps are sets of alternative requirements we can impose on the final distributions of interest; that is, probability gaps are sets of probability sets. We call the elements of a probability gap a *probability insert*. The logic of a probability gap model is thus: we take a probability set representing the *model*, and choose one probability insert from several alternatives that together make a probability gap; the final model of interest is the probability set obtained by intersecting the model with the insert.

This scheme raises the possibility that the model, the insert, or the intersection may be empty if we are not careful. We define the notion of *validity* and show that it is sufficient that the probability model and the probability insert are valid for the intersection to be non-empty.

## 2.6 Probability sets

Note for proofreading

I’ve been adding decorations  $\mathbb{P}$ ,  $\mathbb{P}_{\{\}}$  and  $\mathbb{P}_{\square}$  to denote probability models, probability sets and probability gap models respectively, but I’m not close to adding them everywhere they are needed yet.

end note

A probability set is a set of 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 2.7** (Probability space). A probability space is a triple  $(\mu, \Omega, \mathcal{F})$ , where  $\mu$  is a base measure on  $\mathcal{F}$  which we here take to be  $\mathcal{P}(\Omega)$ .

**Definition 2.8** (Probability set). A probability set  $\mathbb{P}_{\{\}}$  on  $(\Omega, \mathcal{F})$  is a collection of probability measures on  $(\Omega, \mathcal{F})$ . In other words it is a subset of  $\mathcal{P}(\Delta(\Omega))$ .

**Definition 2.9** (Probability gap model). Given a fundamental probability set  $\Omega$  and a set of probability sets  $A$ , a probability gap model  $\mathbb{P}_{\square} : A \rightarrow \mathcal{P}(\Delta(\Omega))$  is a function that sends an element of  $A$  to a probability set on  $\Omega$ .

We will make a substantial simplifying assumption: all sets, including the fundamental probability set  $\Omega$  and any set of values a variable takes, are discrete sets. That is, they are at most countably infinite and the  $\sigma$ -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:  $\mathbb{P}^X(x) := \mathbb{P}^X(\{x\})$ .

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

**Definition 2.10** (Marginal distribution with respect to a probability space). Given a probability space  $(\mu, \Omega)$  and a random variable  $X : \Omega \rightarrow X$ , we can define the *marginal distribution* of  $X$  with respect to  $\mu$ ,  $\mu^X : \mathcal{X} \rightarrow [0, 1]$  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

$$\mu^X = \mu \mathbb{F}_X \quad (41)$$

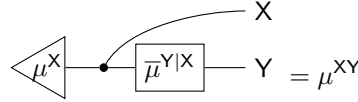
A  $Y|X$ -conditional probability is “the probability of  $Y$  given  $X$  relative to  $\mu$ ”. It is a Markov kernel that maps the marginal distribution of  $X$  to the marginal distribution of the sequence  $(X, Y)$ .

**Definition 2.11** (Conditional probability with respect to a probability space). Given a probability space  $(\mu, \Omega)$  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

$$\mu^X(x)\mu^{Y|X}(y|x) = \mu^{XY}(x, y) \quad \forall x \in X, y \in Y \quad (42)$$

$$\iff \quad (43)$$



$$\quad (44)$$

Given a probability set  $\mathbb{P}_{\{\}}^{\cdot}$ , we define marginal and conditional probabilities as probability measures and Markov kernels that satisfy Definitions 2.10 and 2.11 respectively for *all* base measures in  $\mathbb{P}_{\{\}}$ .

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

**Definition 2.12** (Marginal probability with respect to a probability set). Given a fundamental probability set  $\Omega$  a variable  $X : \Omega \rightarrow X$  and a probability set  $\mathbb{P}_{\{\}}$ , the marginal distribution  $\mathbb{P}_{\{\}}^X = \mathbb{P}_a^X$  for any  $\mathbb{P}_a \in \mathbb{P}_{\{\}}$  if a distribution satisfying this condition exists. Otherwise, it is undefined.

**Definition 2.13** (Conditional probability with respect to a probability set). Given a fundamental probability set  $\Omega$  variables  $X : \Omega \rightarrow X$  and  $Y : \Omega \rightarrow Y$  and a probability set  $\mathbb{P}_{\{\}}$ ,  $\mathbb{P}_{\{\}}^{Y|X}$  is any Markov kernel  $X \rightarrow Y$  such that for all  $\mathbb{P}_a \in \mathbb{P}_{\{\}}$ :

$$\mathbb{P}_a^X(x)\mathbb{P}_{\{\}}^{Y|X}(y|x) = \mathbb{P}_a^{XY}(x, y) \quad \forall x \in X, y \in Y \quad (45)$$

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 it forward”.

**Lemma 2.14** (Equivalence of pushforward definitions). *Given a probability space  $\mathbb{M} : W \rightarrow \Omega$  and  $X : \Omega \rightarrow X$ , define  $\mathbb{K}^{X|W} : W \rightarrow X$  by  $\mathbb{K}^{X|W}(x|w) := \mathbb{M}(X \bowtie x|w)$  for any  $x \in X$  and  $w \in W$  and  $\mathbb{L}^X : W \rightarrow X$  by*

$$\mathbb{L}^{X|W} = \mathbb{M}\mathbb{F}_X \quad (46)$$

Then

$$\mathbb{L}^{X|W} = \mathbb{K}^{X|W} \quad (47)$$

*Proof.* For any  $x \in X$ ,  $w \in W$

$$\mathbb{L}^{X|W}(x|w) = \sum_{\omega \in \Omega} \llbracket x = X(\omega) \rrbracket \mathbb{M}(\omega|w) \quad (48)$$

$$= \sum_{\omega \in X^{-1}(x)} \mathbb{M}(\omega|w) \quad (49)$$

$$= \mathbb{M}(X \bowtie x|w) \quad (50)$$

$$= \mathbb{K}^{X|W}(x|w) \quad (51)$$

□

**Theorem 2.15** (Recursive pushforward). *Suppose we have a fundamental probability set  $\Omega$  variables  $X : \Omega \rightarrow X$  and  $Y : \Omega \rightarrow Y$ ,  $Z : \Omega \rightarrow Z$  and a probability set  $\mathbb{P}_{\{\}}^{\cdot}$  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  $\mathbb{P}_{\{\}}$  given by  $\mathbb{P}_{\{\}}^{Z|X} = \mathbb{P}_{\{\}}^{Y|X} \mathbb{F}_f$ .*

*Proof.* For any  $\mathbb{P}_a \in \mathbb{P}_{\{\}}$ ,  $x, z$

$$\mathbb{P}_a^X(x) \mathbb{P}_a^{Z|X}(z|x) = \mathbb{P}_a(X^{-1}(x) \cap Z^{-1}(z)) \quad (52)$$

$$= \mathbb{P}_a(X^{-1}(x) \cap Y^{-1}(f^{-1}(z))) \quad (53)$$

$$= \mathbb{P}_a^{X,Y}(\{x\} \times f^{-1}(z)) \quad (54)$$

$$= \mathbb{P}_a^X(x) \mathbb{P}_a^{Y|X}(f^{-1}(z)|x) \quad (55)$$

□

We also generalise the notion of almost sure equality with respect to a probability set to mean almost sure equality with respect to every base measure in the set

**Definition 2.16** (Almost sure equality). Two functions  $f : \Omega \rightarrow X$  and  $g : \Omega \rightarrow Y$  are almost surely equal with respect to a probability set  $\mathbb{P}_{\{\}}$  if they are almost surely equal with respect to every base measure  $\mathbb{P} \in \mathbb{P}_{\{\}}$ .

## 2.7 Probability sets defined by marginal and conditional probabilities

In the previous section we defined marginal and conditional probabilities for probability sets. We can define probability sets by marginal or conditional probabilities as the set of all probability models that share those marginals or conditionals.

Not all probability measures  $\mathbb{Q}^X$  on  $X$  define nonempty sets of probability measures on  $\Omega$ . Consider, for example,  $X = (Y, Y)$  for some  $Y : \Omega \rightarrow Y$  and any measure  $\mathbb{Q}^{YY}$  that assigns nonzero probability to the event  $(Y, Y) \bowtie (y, y')$  for  $y \neq y'$ . Then there is no base measure that pushes forward to  $\mathbb{P}^{YY}$ . A *valid*



*candidate distribution* is a distribution associated with a particular variable that defines a nonempty set of base measures on  $\Omega$ .

**Definition 2.17** (Valid candidate distribution). A valid  $\mathbf{X}$ -probability distribution  $\mathbb{P}^{\mathbf{X}}$  is any probability measure on  $\Delta(X)$  such that  $\mathbf{X}^{-1}(A) = \emptyset \implies \mathbb{P}^{\mathbf{X}}(A) = 0$  for all  $A \in \mathcal{X}$ .

**Theorem 2.18** (Validity). *Given  $(\Omega, \mathcal{F})$ ,  $\mathbf{X} : \Omega \rightarrow X$ ,  $\mathbb{J} \in \Delta(X)$  with  $\Omega$  and  $X$  standard measurable, there exists some  $\mu \in \Delta(\Omega)$  such that  $\mu^{\mathbf{X}} = \mathbb{J}$  if and only if  $\mathbb{J}$  is a valid candidate distribution.*

*Proof.* If: This is a Theorem 2.5 of Ershov (1975). Only if: This is also found in Ershov (1975), but is simple enough to reproduce here. Suppose  $\mathbb{J}$  is not a valid probability distribution. Then there is some  $x \in X$  such that  $\mathbf{X} \bowtie x = \emptyset$  but  $\mathbb{J}(x) > 0$ . Then

$$\mu^{\mathbf{X}}(x) = \mu(\mathbf{X} \bowtie x) \quad (56)$$

$$= \sum_{x' \in X} \mathbb{J}(x') \mathbb{K}(\mathbf{X} \bowtie x | x') \quad (57)$$

$$= 0 \quad (58)$$

$$\neq \mathbb{J}(x) \quad (59)$$

□

**Definition 2.19** (Valid candidate conditional). Given  $(\Omega, \mathcal{F})$ ,  $\mathbf{X} : \Omega \rightarrow X$ ,  $\mathbf{Y} : \Omega \rightarrow Y$  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:

$$\forall B \in \mathcal{Y}, x \in \mathcal{X} : (\mathbf{X}, \mathbf{Y}) \bowtie \{x\} \times B = \emptyset \implies \left( \mathbb{P}^{\mathbf{Y}|\mathbf{X}}(A|x) = 0 \right) \vee (\mathbf{X} \bowtie \{x\} = \emptyset) \quad (60)$$

The definition of conditional probability (Definition 2.11) allows in some cases for conditional probabilities that are not valid candidate conditionals. In Theorem 2.29 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.

The particular property we show for valid candidate conditionals is that if we combine them with other valid candidate conditionals via the operation  $\odot$  (Definition 2.20) then we are left with a valid candidate conditional (Theorem 2.25). This reduces to a valid candidate distribution in the case of conditioning on the trivial variable  $*$  (Lemma 2.22). Because  $\odot$  leaves us with a subset of the intersection of the probability sets defined by the two inputs (Theorem 2.21), this also shows that valid candidate conditionals define non-empty probability sets.

Therefore, as long as we limit ourselves to valid candidate conditionals, we can use them to represent probability sets and intersect the sets using  $\odot$  and we are guaranteed not to end up with empty probability sets.

**Definition 2.20** (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

$$\mathbb{K} \odot \mathbb{L} := \text{copy}_X(\mathbb{K} \otimes \text{id}_X)(\text{copy}_Y \otimes \text{id}_X)(\text{id}_Y \otimes \mathbb{L}) \quad (61)$$

$$= \begin{array}{c} \text{Y} \\ \text{X} \text{---} \bullet \text{---} \boxed{\mathbb{K}} \text{---} \bullet \text{---} \boxed{\mathbb{L}} \text{---} \text{Z} \\ \text{---} \text{---} \end{array} \quad (62)$$

$$\iff \quad (63)$$

$$(\mathbb{K} \odot \mathbb{L})(A \times B|x) = \int_B \mathbb{L}(A|y, x) \mathbb{K}(dy|x) \quad (64)$$

**Theorem 2.21** (Copy-product is an intersection of probability sets). *Given  $\Omega, X, Y, Z$  all standard measurable, and valid candidate conditionals  $\mathbb{P}_{\{\}}^{Y|X}$  and  $\mathbb{Q}_{\{\}}^{Z|YX}$  with  $\mathbb{P}_{\{\}}$  the largest probability set with conditional  $\mathbb{P}_{\{\}}^{Y|X}$  and  $\mathbb{Q}_{\{\}}$  the largest probability set with conditional  $\mathbb{Q}_{\{\}}^{Z|YX}$ , then the largest probability set  $\mathbb{R}_{\{\}}$  with conditional  $\mathbb{R}_{\{\}}^{YZ|X} = \mathbb{P}_{\{\}}^{Y|X} \odot \mathbb{Q}_{\{\}}^{Z|YX}$  is equal to  $\mathbb{R}_{\{\}} = \mathbb{P}_{\{\}} \cap \mathbb{Q}_{\{\}}$ .*

Make sure this works for a.s. equality

*Proof.* By Theorem 2.15,  $\mathbb{R}_{\{\}}^{Y|X} = \mathbb{P}_{\{\}}^{Y|X}$  so  $\mathbb{R}_{\{\}} \subset \mathbb{P}_{\{\}}$ . Furthermore by construction  $\mathbb{R}_{\{\}}^{Z|YX} = \mathbb{Q}_{\{\}}^{Z|YX}$  so  $\mathbb{R}_{\{\}} \subset \mathbb{P}_{\{\}} \cap \mathbb{Q}_{\{\}}$ .

Suppose there's an element  $\mathbb{S}$  of  $\mathbb{P}_{\{\}} \cap \mathbb{Q}_{\{\}}$  not in  $\mathbb{R}_{\{\}}$ . Then  $\mathbb{S}^{YZ|X} \neq \mathbb{R}_{\{\}}^{YZ|X}$ . But by assumption  $\mathbb{S}^{Z|YX} = \mathbb{R}_{\{\}}^{Z|YX}$  and  $\mathbb{S}^{Y|X} = \mathbb{R}_{\{\}}^{Y|X}$ . But then  $\mathbb{S}^{YX|X} = \mathbb{R}_{\{\}}^{Y|X} \odot \mathbb{R}_{\{\}}^{Z|YX} = \mathbb{R}_{\{\}}^{YZ|X}$ , contradiction.  $\square$

**Lemma 2.22** (Equivalence of validity definitions). *Given  $X : \Omega \rightarrow X$ , with  $\Omega$  and  $X$  standard measurable, 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

$$\forall A \in \mathcal{X} : X \bowtie A = \emptyset \implies \mathbb{P}^{X|*}(A|*) = 0 \quad (65)$$

But  $\mathbb{P}^{X|*}(A|*) = \mathbb{P}^X(A)$  by definition, so this is equivalent to

$$\forall A \in \mathcal{X} : X \bowtie A = \emptyset \implies \mathbb{P}^X(A) = 0 \quad (66)$$

$\square$

**Lemma 2.23** (Copy-product of valid candidate conditionals is valid). *Given  $(\Omega, \mathcal{F})$ ,  $X : \Omega \rightarrow X$ ,  $Y : \Omega \rightarrow Y$ ,  $Z : \Omega \rightarrow Z$  (all spaces standard measurable) 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 for each  $x \in X(\Omega)$ , as it is automatically satisfied for other values of  $X$ .

For all  $x \in X(\Omega)$ ,  $B \in \mathcal{Y}$  such that  $X \bowtie \{x\} \cap Y \bowtie B = \emptyset$ ,  $\mathbb{P}^{Y|X}(B|x) = 0$  by validity. Thus for arbitrary  $C \in \mathcal{Z}$

$$\mathbb{R}^{YZ|X}(B \times C|x) = \int_B \mathbb{Q}^{Z|YX}(C|y, x) \mathbb{P}^{Y|X}(dy|x) \quad (67)$$

$$\leq \mathbb{P}^{Y|X}(B|x) \quad (68)$$

$$= 0 \quad (69)$$

For all  $\{x\} \times B$  such that  $X \bowtie \{x\} \cap Y \bowtie B \neq \emptyset$  and  $C \in \mathcal{Z}$  such that  $(X, Y, Z) \bowtie \{x\} \times B \times C = \emptyset$ ,  $\mathbb{Q}^{Z|YX}(C|y, x) = 0$  for all  $y \in B$  by validity. Thus:

$$\mathbb{R}^{YZ|X}(B \times C|x) = \int_B \mathbb{Q}^{Z|YX}(C|y, x) \mathbb{P}^{Y|X}(dy|x) \quad (70)$$

$$= 0 \quad (71)$$

□

**Corollary 2.24** (Valid candidate conditional is validly extendable to a valid candidate distribution). *Given  $\Omega$ ,  $U : \Omega \rightarrow U$ ,  $W : \Omega \rightarrow W$  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 2.23 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 2.22. □

**Theorem 2.25** (Validity of conditional probabilities). *Suppose we have  $\Omega$ ,  $X : \Omega \rightarrow X$ ,  $Y : \Omega \rightarrow Y$ , with  $\Omega$ ,  $X$ ,  $Y$  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 2.24.

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

$$(\mathbb{V}^X \odot \mathbb{T}^{Y|X})(x, y) = \mathbb{T}^{Y|X}(y|x) \mathbb{V}^X(x) \quad (72)$$

$$= \mathbb{T}^{Y|X}(y|x) \quad (73)$$

$$> 0 \quad (74)$$

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

## 2.8 Probability gap models

I need a statement like: probability gap model of type  $X|Y$  with insert  $Z|X$

A probability gap model is a probability set equipped with a domain, which is a set of probability sets and we refer to it as a *probability gap*. Here we focus on probability sets that can be intersected using the operation  $\odot$  (see Theorem 2.21). Thus we can equivalently view probability gap models as “incomplete sequences” of conditional probabilities, and probability gaps as sets of alternative conditional probabilities that complete the sequences.

**A note on terminology:** Probability gap models are written with blackboard letters  $\mathbb{P}_{\square}$ . The same base letter with different superscripts  $\mathbb{P}_{\square}^{A|B}$  indicates a conditional probability with respect to  $\mathbb{P}_{\square}$ . We use subscripts to indicate both the model obtained by applying particular choices of insert  $\mathbb{P}_{\alpha} := \mathbb{P}_{\square} \cap \alpha$ .  $A \perp_{\mathbb{P}_{\square}} B$  is a statement of independence with respect to the model  $\mathbb{P}_{\square}$ .

The important features of probability gap models are their types – what kind of probability sets they accept, and what kind of probability set we end up with as a result. We will limit our attention to probability sets defined by conditional probabilities, so that the intersection operation corresponds to the product of conditional probabilities via  $\odot$ . The general form of such models are collections of conditional probabilities with some conditional probabilities missing – the set of conditional probabilities that could fill in the missing parts is the probability gap.

**Definition 2.26** (Marginal probability gap model). Given probability set  $\Omega$  and variables  $X : \Omega \rightarrow X$  and  $Y : \Omega \rightarrow Y$  and a set  $A$  of valid candidate distributions on  $X$ , a probability gap model with respect to  $X$ , written  $\mathbb{P}_{\square} : A \rightarrow \mathcal{P}(\Delta(\Omega))$ , is map from  $A$  to probability sets on  $\Omega$  and is defined by a valid candidate conditional  $\mathbb{P}_{\square}^{Y|X}$  and implements the map

$$\alpha \mapsto \alpha \odot \mathbb{P}_{\square}^{Y|X} \quad (75)$$

$$=: \mathbb{P}_{\alpha}^{XY} \quad (76)$$

For all  $\alpha \in A$ .

We can think of a marginal probability gap model as a probability model “ $\mathbb{P}$ ” that is missing the marginal probability “ $\mathbb{P}^X$ ”. We can insert any of a number of different options  $\mathbb{P}_{\alpha}^X$  for each  $\alpha \in A$ , an operation that yields the probability set defined by  $\mathbb{P}_{\alpha}^{XY}$  (which, if the sample space is just  $X \times Y$ , is just a probability model).

Applying the same idea, but with a missing conditional  $\mathbb{P}_{\alpha}^{Y|X}$  yields what we call a conditional probability gap model. A conditional gap model  $\mathbb{P}_{\square}$  is defined by  $(\mathbb{P}_{\square}^X, \mathbb{P}_{\square}^{Z|YX}, A)$ . Note that  $\mathbb{P}_{\square}^X$  and  $\mathbb{P}_{\square}^{Z|YX}$  are of the wrong type to combine using  $\odot$ ; we need a conditional probability  $\mathbb{P}_{\alpha}^{Y|X}$  to fill the “gap” between them.

**Definition 2.27** (Conditional probability gap model). Given probability set  $\Omega$  and variables  $X : \Omega \rightarrow X$ ,  $Y : \Omega \rightarrow Y$ ,  $Z : \Omega \rightarrow Z$  and a set  $A$  of valid candidate

conditionals on  $X \rightarrow Y$ , an order 2 probability gap model  $\mathbb{P}_\square : A \rightarrow \mathcal{P}(\Delta(\Omega))$  is map from  $A$  to probability sets on  $\Omega$ , is defined by a valid pair  $(\mathbb{P}_\square^X, \mathbb{P}_\square^{Z|XY})$  and implements the map

$$\alpha \mapsto (\mathbb{P}_\square^X \odot \alpha) \odot \mathbb{P}_\square^{Z|XY} \quad (77)$$

$$=: \mathbb{P}_\alpha^{XYZ} \quad (78)$$

for all  $\alpha \in A$ .

A pair of conditional probabilities with a gap between them induces a *probability 2-combs* (Chiribella et al., 2008; Jacobs et al., 2019). We can depict the map associated with a conditional gap model graphically in an informal way as “inserting”  $\mathbb{P}_\alpha^{Y|X}$  into  $\mathbb{P}_\square^{Z|XY}$ :

$$\text{Insert} \left( \begin{array}{c} \left( \begin{array}{c} \triangleleft \mathbb{P}_\square^X \end{array} \begin{array}{c} \bullet \\ \text{---} X \end{array} \begin{array}{c} Y \end{array} \begin{array}{c} \square \mathbb{P}_\square^{Z|XY} \end{array} \text{---} Z \end{array} \right) , \\ \begin{array}{c} X \text{---} \square \mathbb{P}_\alpha^{Y|XW} \text{---} X \end{array} \end{array} \right) \quad (79)$$

$$= \begin{array}{c} \triangleleft \mathbb{P}_\square^X \bullet \begin{array}{c} \square \mathbb{P}_\alpha^{Y|X} \end{array} \bullet \begin{array}{c} \square \mathbb{P}_\square^{Z|WXY} \end{array} \begin{array}{c} \text{---} Z \\ \text{---} Y \\ \text{---} X \end{array} \end{array} \quad (80)$$

We can generalise the idea of conditional probabilities with gaps between them. For example, we could have a probability gap model defined by a triple of conditional probabilities  $(\mathbb{P}_\square^{X_1}, \mathbb{P}_\square^{X_3|X_1X_2}, \mathbb{P}_\square^{X_5|X_1X_2X_3X_4}, A)$  with inserts of type  $(\mathbb{P}_\alpha^{X_1}, \mathbb{P}_\alpha^{X_4|X_1X_2X_3})$ , and the map is given by selecting conditionals alternately from the model and the insert recursively applying  $\odot$

$$\alpha \mapsto (((\mathbb{P}_\square^{X_1} \odot \mathbb{P}_\alpha^{X_1}) \quad (81)$$

$$\odot \mathbb{P}_\square^{X_3|X_1X_2}) \odot \mathbb{P}_\alpha^{X_4|X_1X_2X_3}) \quad (82)$$

$$\odot \mathbb{P}_\square^{X_5|X_1X_2X_3X_4} \quad (83)$$

History-based models of reinforcement found in Hutter (2004) are examples of probability gap models of this type. Such models posit sequences of variables  $(A_i, O_i, R_i)_{i \in \mathbb{N}}$  representing the  $i$ th action, observation and reward respectively. They also posit an *agent* that implements a *policy*, which is a collection of maps  $\pi_i : O^i \times R^i \times A^i \rightarrow A$  for all  $i \in \mathbb{N}$  (mapping the history of actions, rewards and observations to the next action), and an *environment* which is a collection of maps  $e_i : O^i \times R^i \times A^{i+1} \rightarrow O \times R$ , mapping the history of actions, rewards, observations and the next action to the next observation and reward.

We can identify the environment  $e_i$  with a probability gap model  $\{\mathbb{P}_{\square}^{O_i R_i | O <_i R <_i A_i} | i \in \mathbb{N}\}$  and the policy set with a probability gap  $\{\mathbb{P}_{\alpha}^{A_i | O <_i R <_i A_i} | i \in \mathbb{N}, \alpha \in A\}$ ; the model of observations, actions and rewards arising from a particular choice  $\alpha$  of policy is determined by the composition of environment and policy conditional probabilities as in Equation 83.

We will define a shorthand for models of this type. A probability  $n$ -comb is a Markov kernel  $\mathbb{K}_n : \prod_{i \in [n-1]} A_i \rightarrow \prod_{i \in [n-1]} B_i$  such that there is an  $n - 1$  comb  $\mathbb{K}_{n-1} : \prod_{i \in [n-1]} A_i \rightarrow \prod_{i \in [n-1]} B_i$  satisfying

$$\begin{array}{c} A_{[n-1]} \\ A_n \end{array} \begin{array}{c} \text{---} \\ \text{---} \end{array} \boxed{\mathbb{K}_n} \begin{array}{c} \text{---} \\ \text{---} \end{array} \begin{array}{c} B_{[n-1]} \\ * \end{array} = \begin{array}{c} A_{[n-1]} \\ A_n \end{array} \begin{array}{c} \text{---} \\ \text{---} \end{array} \boxed{\mathbb{K}_{n-1}} \begin{array}{c} \text{---} \\ \text{---} \end{array} \begin{array}{c} B_{[n-1]} \\ * \end{array} \quad (84)$$

And a  $1$ -comb is simply an arbitrary Markov kernel  $\mathbb{K}_1 : A_1 \rightarrow B_1$ . A probability  $\mathbb{N}$ -comb is a collection of  $n$ -combs  $\mathbb{K}_n$  for each  $n \in \mathbb{N}$  such that  $\mathbb{K}_{n-1}$  is related to  $\mathbb{K}_n$  by Equation 84.

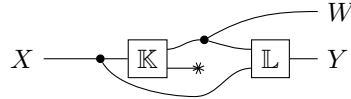
We can specify a probability gap model  $\mathbb{P}_{\square}$  associated with an infinite collection of conditionals  $\{\mathbb{P}_{\square}^{X_{2i+1} | X_i} | i \in \mathbb{N}\}$  with an  $\mathbb{N}$ -comb  $\{\mathbb{K}_i : X^i \rightarrow X^i | i \in \mathbb{N}\}$ . Crucially, we can specify the *signature* of this  $\mathbb{N}$ -comb in terms of which variables it takes as inputs and which it maps as outputs; in particular, it takes  $(X_{2i})_{i \in \mathbb{N}}$  as input and  $(X_{2i+1})_{i \in \mathbb{N}}$  as output. Thus we can specify a *probability sequence gap model* with an  $\mathbb{N}$ -comb, which we will write as  $\mathbb{P}_{\square}^{X_{2N+1} || X_{2N}}$ , noting the double bar  $||$  to indicate that this is a comb rather than a conditional probability. This is in turn associated with probability gaps that are subsets of the set of  $\mathbb{N}$ -combs  $\{\mathbb{P}_{\alpha}^{X_{2N} | X_{2N+1}} | \alpha \in A\}$ .

It's an open question whether we can represent the comb  $\mathbb{P}_{\square}^{X_{2N+1} || X_{2N}}$  with a single Markov kernel  $\mathbb{N} \rightarrow \mathbb{N}$  in the vein of Kolmogorov's representation theorem for probability measures.

### 2.8.1 Disintegrations

Markov kernels on standard measurable spaces can be disintegrated into a pairs of kernels that yield the original kernel when composed with  $\odot$ . We can find conditional probabilities using disintegrations.

**Lemma 2.28** (Disintegration existence in discrete Markov kernels). *For any Markov kernel  $\mathbb{K} : X \rightarrow W \times Y$  and  $X, W, Y$  are standard measurable, there exists  $\mathbb{L} : W \times X \rightarrow Y$  such that*



$$\mathbb{K} = \quad (85)$$

$\mathbb{L}$  is a disintegration of  $\mathbb{K}$ .

*Proof.* Cho and Jacobs (2019) Theorem 3.11

□

We can

**Theorem 2.29** (Existence of conditional probabilities). *Given a probability gap model  $\mathbb{P}_\square : A \rightarrow \Delta(\Omega)$  along with a valid conditional probability  $\mathbb{P}_\square^{XY|W}$ , there exists a valid conditional probability  $\mathbb{P}_\square^{Y|WX}$ .*

*Proof.* From Lemma 2.28, we have the existence of some Markov kernel  $\mathbb{P}_\square^{Y|WX} : W \times X \rightarrow Y$  such that

$$\mathbb{P}_\square^{XY|W} = \mathbb{P}_\square^{X|W} \odot \mathbb{P}_\square^{Y|WX} \quad (86)$$

By definition of conditional probability, for any insert  $\alpha \in A$  there exists  $\mathbb{P}_\alpha^W \in \Delta(W)$  such that

$$\mathbb{P}_\alpha^{WXY} = \mathbb{P}_\alpha^W \odot \mathbb{P}_\square^{XY|W} \quad (87)$$

Thus

$$\mathbb{P}_\alpha^{WXY} = \mathbb{P}_\alpha^W \odot (\mathbb{P}_\square^{X|W} \odot \mathbb{P}_\square^{Y|WX}) \quad (88)$$

$$= (\mathbb{P}_\alpha^W \odot \mathbb{P}_\square^{X|W}) \odot \mathbb{P}_\square^{Y|WX} \quad (89)$$

Let  $\text{erase}_Y : Y \rightarrow \{*\}$  be the erase function on  $Y$  (as opposed to the erase kernel) and  $\text{idf}_{W \times X}$  be the identity function on  $W \times X$ . Noting that

$$(W, X) = (\text{idf}_{W \times X} \otimes \text{erase}_Y) \circ (W, X, Y) \quad (90)$$

By Lemma 2.14 together with Theorem 2.15 we have for all  $\alpha$ :

$$\mathbb{P}_\alpha^{XW} = \mathbb{P}_\alpha^{WXY} (\text{id}_{W \times X} \otimes \text{erase}_Y) \quad (91)$$

$$= \mathbb{P}_\alpha^W \odot (\mathbb{P}_\square^{X|W} \odot \mathbb{P}_\square^{Y|WX}) (\text{id}_{W \times X} \otimes \text{erase}_Y) \quad (92)$$

$$= \mathbb{P}_\alpha^W \odot \mathbb{P}_\square^{X|W} \quad (93)$$

Then

$$\mathbb{P}_\alpha^{XWY} = (\mathbb{P}_\alpha^{XW}) \odot \mathbb{P}_\square^{Y|WX} \quad (94)$$

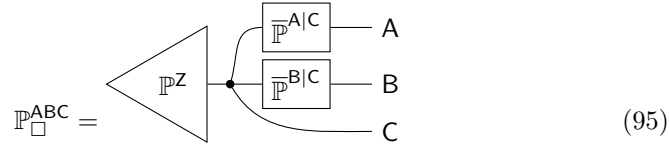
And so  $\mathbb{P}_\square^{Y|WX}$  is a  $Y|WX$  conditional probability. We also want it to be valid, so we will verify that it can be chosen as such.

We also need to check that  $\mathbb{P}_\square^{Y|WX}$  can be chosen so that it is valid. By validity of  $\mathbb{K}^{W,Y|X}$ ,  $w \in W(\Omega)$  and  $(X, W, Y) \bowtie (x, w, y) = \emptyset \implies \mathbb{P}_\square^{W,Y|X} = 0$ , so we only need to check for  $(w, x, y)$  such that  $\mathbb{P}_\square^{W,Y|X}(w, y|x) = 0$ . For all  $x, y$  such that  $\mathbb{K}^{Y|X}(y|x)$  is positive, we have  $\mathbb{P}_\square^{W,Y|X}(w, y|x) = 0 \implies \mathbb{P}_\square^{Y|WX}(y|w, x) = 0$ . 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  $\mathbb{P}_\square^{Y|WX}(Y(\omega)|w, x) = 1$ . □

### 2.8.2 Conditional independence

Conditional independence has a familiar definition in probability models. We define conditional independence with respect to a probability gap model to be equivalent to conditional independence with respect to every base measure in the range of the model. This definition is closely related to the idea of *extended conditional independence* proposed by Constantinou and Dawid (2017).

**Definition 2.30** (Conditional independence with respect to a probability model). For a *probability model*  $\mathbb{P}_\square$  and variables  $A, B, Z$ , we say  $B$  is conditionally independent of  $A$  given  $C$ , written  $B \perp\!\!\!\perp_{\mathbb{P}_\square} A|C$ , if

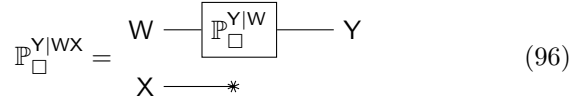


$$\mathbb{P}_\square^{ABC} = \quad (95)$$

For any  $\mathbb{P}_\square^{B|C}$  and  $\mathbb{P}_\square^{A|C}$ . 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:  $A \perp\!\!\!\perp_{\mathbb{P}_\square} B|C$  iff  $B \perp\!\!\!\perp_{\mathbb{P}_\square} A|C$
2. Decomposition:  $A \perp\!\!\!\perp_{\mathbb{P}_\square} (B, C)|W$  implies  $A \perp\!\!\!\perp_{\mathbb{P}_\square} B|W$  and  $A \perp\!\!\!\perp_{\mathbb{P}_\square} C|W$
3. Weak union:  $A \perp\!\!\!\perp_{\mathbb{P}_\square} (B, C)|W$  implies  $A \perp\!\!\!\perp_{\mathbb{P}_\square} B|(C, W)$
4. Contraction:  $A \perp\!\!\!\perp_{\mathbb{P}_\square} C|W$  and  $A \perp\!\!\!\perp_{\mathbb{P}_\square} B|(C, W)$  implies  $A \perp\!\!\!\perp_{\mathbb{P}_\square} (B, C)|W$

**Theorem 2.31.** Given discrete  $\Omega$  and a probability model  $\mathbb{P}_\square$  and variables  $W : \Omega \rightarrow W$ ,  $X : \Omega \rightarrow X$ ,  $Y : \Omega \rightarrow Y$ ,  $Y \perp\!\!\!\perp_{\mathbb{P}_\square} X|W$  if and only if there exists some version of  $\mathbb{P}_\square^{Y|WX}$  and  $\mathbb{P}_\square^{Y|W}$  such that



$$\mathbb{P}_\square^{Y|WX} = \frac{W \text{ --- } \boxed{\mathbb{P}_\square^{Y|W}} \text{ --- } Y}{X \text{ --- } *} \quad (96)$$

$$\iff \mathbb{P}_\square^{Y|WX}(y|w, x) = \mathbb{P}_\square^{Y|W}(y|w) \quad (97)$$

*Proof.* See Cho and Jacobs (2019).  $\square$

**Definition 2.32** (Conditional independence with respect to a probability gap model). Conditional independence  $A \perp\!\!\!\perp_{\mathbb{P}_\square} B|C$  holds for an arbitrary probability gap model  $\mathbb{P}_\square : A \rightarrow \mathcal{P}(\Delta(\Omega))$  if  $A \perp\!\!\!\perp_{\mathbb{P}_\alpha} B|C$  holds for all probability models  $\mathbb{P}_\alpha$ ,  $\alpha \in A$ .

One case where we can deduce conditional independences in probability gap models is when conditional probabilities exist and they are *unresponsive* to some input variables.



**Definition 2.33** (Unresponsiveness). Given discrete  $\Omega$ , a probability gap model  $\mathbb{P}_\square : A \rightarrow \Delta(\Omega)$ , variables  $W : \Omega \rightarrow W$ ,  $X : \Omega \rightarrow X$ ,  $Y : \Omega \rightarrow Y$ , if there is some version of the conditional probability  $\mathbb{P}^{Y|WX}$  and  $\mathbb{P}_\square^{Y|W}$  such that

$$\mathbb{P}_\square^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{P}_\square^{Y|W}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (98)$$

then  $\mathbb{P}_\square^{Y|WX}$  is *unresponsive* to  $X$ .

**Definition 2.34** (Domination). Given a probability gap model  $\mathbb{P}_\square : A \rightarrow \Delta(\Omega)$ ,  $\alpha \in A$  dominates  $A$  if  $\mathbb{P}_\beta(B) > 0 \implies \mathbb{P}_\alpha(B) > 0$  for all  $\beta \in A$ ,  $B \in \mathcal{F}$ .

**Theorem 2.35** (Conditional independence from kernel unresponsiveness). *Given discrete  $\Omega$ , variables  $W : \Omega \rightarrow W$ ,  $X : \Omega \rightarrow X$ ,  $Y : \Omega \rightarrow Y$  and a probability gap model  $\mathbb{P}_\square : A \rightarrow \Delta(\Omega)$  with conditional probability  $\mathbb{P}_\square^{Y|WX}$  and such that there is  $\alpha \in A$  dominating  $A$ ,  $Y \perp\!\!\!\perp_{\mathbb{P}_\square} X|W$  if and only if  $\mathbb{P}_\square^{Y|WX}$  is unresponsive to  $W$ .*

*Proof.* If: For every  $\alpha \in A$  we can write

$$\mathbb{P}_\alpha^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{P}_\alpha^{Y|W}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (99)$$

And so, by Theorem 2.31,  $Y \perp\!\!\!\perp_{\mathbb{P}_\alpha} X|W$  for all  $\alpha \in A$ , and so  $Y \perp\!\!\!\perp_{\mathbb{P}_\square} X|W$ . Only if: By Theorem 2.31, there exists a version of  $\mathbb{P}_\alpha^{Y|WX}$  unresponsive to  $W$ . Because  $\alpha$  dominates  $A$ , every version of  $\mathbb{P}_\alpha^{Y|WX}$  is a version of  $\mathbb{P}_\beta^{Y|WX}$  for all  $\beta \in A$ , thus it is a version of  $\mathbb{P}_\square^{Y|WX}$  also.  $\square$

Note that  $Y \perp\!\!\!\perp_{\mathbb{P}_\square} X|W$  does *not* imply the existence of  $\mathbb{P}_\square^{Y|WX}$ . If we have, for example,  $A = \{\alpha, \beta\}$  and  $\mathbb{P}_\alpha^{AB}$  is two flips of a fair coin while  $\mathbb{P}_\beta^{AB}$  is a flip of a biased coin followed by a flip of a fair coin, then  $A \perp\!\!\!\perp_{\mathbb{P}} B$  but  $\mathbb{P}^{AB}$  does not exist.

We also need the domination condition. Consider  $A$  a collection of inserts that all deterministically set a variable  $X$ ; then for any variable  $Y$   $Y \perp\!\!\!\perp_{\mathbb{P}_\square} X$  because  $X$  is deterministic for any  $\alpha \in A$ . But  $\mathbb{P}_\square^{Y|X}$  is not necessarily unresponsive to  $X$ .

### 2.8.3 Extended conditional independence

In the case of a probability gap model  $(\mathbb{P}_\square^{Y|W}, A)$  where there is some  $\alpha \in A$  dominating  $A$ , we can relate conditional independence with respect to  $\mathbb{P}_\square$  to what Constantinou and Dawid (2017) *extended conditional independence*, which is a notion they define with respect to a Markov kernel. These concepts may differ if  $A$  is not dominated. Theorem 4.4 of Constantinou and Dawid (2017) proves the following claim:

**Theorem 2.36.** 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  $\mathbb{P}_\alpha^W$  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  $\mathbb{P}_\alpha^{VW} := \mathbb{P}_\alpha^W \odot \mathbb{P}^{V|W}$

$$A \perp\!\!\!\perp_{\mathbb{P}, ext} (B, D)|(C, E) \iff A^* \perp\!\!\!\perp_{\mathbb{P}_\alpha} (B^*, D^*|(C^*, E^*) \quad (100)$$

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

$$A \perp\!\!\!\perp_{\mathbb{P}, ext}^1 (B, D)|(C, E) \iff A^* \perp\!\!\!\perp_{\mathbb{P}} (B^*, D^*|(C^*, E^*) \quad (101)$$

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

Only if:

For any  $\beta$ ,  $\mathbb{P}_\beta^{ABC|DE} = \mathbb{P}_\beta^{DE} \odot \mathbb{P}^{ABC|DE}$ . By Lemma 2.28, we have  $\mathbb{P}^{A|BCDE}$  such that

$$\mathbb{P}_\beta^{A^*B^*C^*D^*E^*} = \mathbb{P}_\beta^{D^*E^*} \odot \mathbb{P}^{B^*C^*|D^*E^*} \odot \mathbb{P}^{A^*|B^*C^*D^*E^*} \quad (102)$$

$$= \mathbb{P}_\beta^{B^*C^*D^*E^*} \odot \mathbb{P}^{A^*|B^*C^*D^*E^*} \quad (103)$$

$$= \mathbb{P}_\beta^{C^*E^*} \odot \mathbb{P}_\beta^{B^*D^*|C^*E^*} \odot \mathbb{P}^{A^*|B^*C^*D^*E^*} \quad (104)$$

By Theorem 2.36, we have some  $\alpha$  such that  $\mathbb{P}_\alpha^{D^*E^*}$  is strictly positive on the range of  $(D^*, E^*)$  and  $A^* \perp\!\!\!\perp_{\mathbb{P}_\alpha} (B^*, D^*|(C^*, E^*)$ .

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

$$\begin{aligned}
\mathbb{P}_\alpha^{C^*E^*} \odot \mathbb{P}_\alpha^{B^*D^*|C^*E^*} \odot \mathbb{P}^{A^*|B^*C^*D^*E^*} &= \text{Diagram (105)} \\
&= \text{Diagram (106)} \\
&= \mathbb{P}_\alpha^{C^*E^*} \odot \mathbb{P}_\alpha^{B^*D^*|C^*E^*} \odot (\mathbb{P}_\alpha^{A^*|C^*E^*} \otimes \text{erase}_{BD}) \quad (107)
\end{aligned}$$

Diagram (105) shows a triangle labeled  $\mathbb{P}_\alpha^{C^*E^*}$  with three outputs:  $A^*$  (via box  $\overline{\mathbb{P}}_\alpha^{A^*|C^*E^*}$ ),  $B^*D^*$  (via box  $\overline{\mathbb{P}}_\alpha^{B^*D^*|C^*E^*}$ ), and  $C^*E^*$ .

Diagram (106) shows the same triangle, but the output  $A^*$  is now via box  $\overline{\mathbb{P}}_\alpha^{A^*|C^*E^*}$  and the output  $B^*D^*$  is via box  $\overline{\mathbb{P}}_\alpha^{B^*D^*|C^*E^*}$ . A star symbol  $*$  is placed on the line connecting the two boxes.

Thus for any  $(a, b, c, d, e) \in A \times B \times C \times D \times E$  such that  $\mathbb{P}_\alpha^{B^*C^*D^*E^*}(b, c, d, e) > 0$ ,  $\mathbb{P}^{A^*|B^*C^*D^*E^*}(a|b, c, d, e) = \mathbb{P}_\alpha^{A^*|C^*E^*}(a|c, e)$ . 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  $\mathbb{P}_\beta^{A^*|B^*C^*D^*E^*} = \mathbb{P}_\alpha^{A^*|C^*E^*}(a|c, e)$  everywhere except a set of  $\mathbb{P}_\beta$ -measure 0. Thus

$$\mathbb{P}_\beta^{A^*B^*C^*D^*E^*} = \text{Diagram (108)}$$

Diagram (108) shows a triangle labeled  $\mathbb{P}_\beta^{C^*E^*}$  with three outputs:  $A^*$  (via box  $\overline{\mathbb{P}}_\alpha^{A^*|C^*E^*}$ ),  $B^*D^*$  (via box  $\overline{\mathbb{P}}_\beta^{B^*D^*|C^*E^*}$ ), and  $C^*E^*$ . A star symbol  $*$  is placed on the line connecting the two boxes.

$$= \text{Diagram (109)}$$

Diagram (109) shows a triangle labeled  $\mathbb{P}_\beta^{C^*E^*}$  with three outputs:  $A^*$  (via box  $\overline{\mathbb{P}}_\alpha^{A^*|C^*E^*}$ ),  $B^*D^*$  (via box  $\overline{\mathbb{P}}_\beta^{B^*D^*|C^*E^*}$ ), and  $C^*E^*$ .

□

### 2.8.4 Graphical properties of conditional independence

It is well-known that directed acyclic graphs are able to represent some conditional independence properties of probability models via the graphical property of *d-separation*. String diagrams are similar to directed acyclic graphs, and string diagrams can be translated into directed acyclic graphs and vice-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 2.35. 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.

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

### 2.9.1 Repeated variables

Lemmas 2.38 and 2.39 establish that models of repeated variables must connect the repetitions with a copy map.

**Lemma 2.38** (Output copies of the same variable are identical). *For any  $\Omega$ ,  $X, Y, Z$  random variables on  $\Omega$  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*

$$X \text{ --- } \boxed{\mathbb{K}^{YYZ|X}} \begin{array}{l} \text{---}^* Y \\ \text{---} Z \end{array} = \mathbb{K}^{YZ|X} \quad (110)$$

and it is given by

$$\mathbb{K}^{YYZ|X} = X \text{ --- } \boxed{\mathbb{K}^{YZ|X}} \begin{array}{l} \text{---} Y \\ \text{---} Y \\ \text{---} Z \end{array} \quad (111)$$

$$\iff \quad (112)$$

$$\mathbb{K}^{YYZ|X}(y, y', z|x) = \llbracket y = y' \rrbracket \mathbb{K}^{YZ|X}(y, z|x) \quad (113)$$

$$(114)$$

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

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

$$\mathbb{K}^{YYZ|X}(y, y', z|x) = \llbracket y = y' \rrbracket \mathbb{L}(y, z|x) \quad (115)$$

$$(116)$$

But then

$$\mathbb{K}^{YZ|X}(y, z|x) = \sum_{y' \in Y} \mathbb{K}^{YYZ|X}(y, y', z|x) \quad (117)$$

$$= \mathbb{L}(y, z|x) \quad (118)$$

$$(119)$$

□

**Lemma 2.39** (Copies shared between input and output are identical).

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

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

$$\mathbb{K} = \begin{array}{c} X \\ Y \end{array} \begin{array}{c} \text{---} \bullet \text{---} \\ \text{---} \end{array} \begin{array}{c} \text{---} X \\ \boxed{\mathbb{K}^{Z|XY}} \\ \text{---} Z \end{array} \quad (120)$$

$$\iff \quad (121)$$

$$\mathbb{K}_{x,y}^{x',z} = \llbracket x = x' \rrbracket \mathbb{L}_{x,y}^z \quad (122)$$

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

$$\begin{array}{c} X \\ Y \end{array} \begin{array}{c} \text{---} \\ \text{---} \end{array} \begin{array}{c} \boxed{\mathbb{K}^{XZ|XY}} \\ \text{---} * \\ \text{---} Z \end{array} = \mathbb{K}^{XZ|XY} \quad (123)$$

and it is given by

$$\mathbb{K}^{XZ|XY} = X \text{---} \begin{array}{c} \boxed{\mathbb{K}^{YZ|X}} \\ \text{---} Y \\ \text{---} Z \end{array} \begin{array}{c} \text{---} Y \\ \bullet \text{---} Y \end{array} \quad (124)$$

$$\iff \quad (125)$$

$$\mathbb{K}^{XZ|XY}(x, z|x', y) = \llbracket x = x' \rrbracket \mathbb{K}^{Z|XY}(z|x', y) \quad (126)$$

$$(127)$$

*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  $\mathbb{L} : X \times Y \rightarrow Z$  such that

$$\mathbb{K}^{XZ|XY}(x', z|x, y) = 0 = \llbracket x = x' \rrbracket \mathbb{L}(z|x, y) \quad (128)$$

$$(129)$$

But then

$$\mathbb{K}^{Z|XY}(y, z|x) = \sum_{x' \in X} \mathbb{K}^{XZ|XY}(x', z|x, y) \quad (130)$$

$$= \mathbb{L}(z|x, y) \quad (131)$$

$$(132)$$

□

start again here

### 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”.

It is an open question whether the models induced by any of these theories are equivalent to probability gap models.

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, 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*  $Q_\alpha^{D|X} : X \rightarrow D$ , letting  $A$  be the set of all such functions available to be chosen. That is, we need a higher order function  $f$  that takes a decision function  $Q_\alpha^{X|D}$  and returns a probabilistic model of the consequences of selecting that decision function  $\mathbb{P}_\alpha^{DXY}$ . An order 2 model  $(\mathbb{P}_\square^{X, \mathbb{P}_\square^Y | XD}, A)$  defines such a function, though there are many such functions that are not order 2 models. The key feature of probability gap models is that the map is by intersection of probability sets, so for example the conditional probability of  $X|D$  given a decision function  $Q_\alpha^{X|D}$  must actually be equal to  $Q_\alpha^{X|D}$ , and we can say the same for  $\mathbb{P}_\square^X$  and  $\mathbb{P}_\square^{Y|XD}$ . If we don't think all of these conditional probabilities are fixed, then we want something other than an order 2 model of the type discussed. We will define *ordinary decision problems* to be those for which the desired model  $\mathbb{P}_\square$  is this type of order 2 probability gap model.

I think adding hypotheses at this point might make things unnecessarily confusing; on the other hand, they are useful for the connection to classical statistical decision theory. The "repeatable experiments" section shows how see-do models with certain assumptions induce an easier to understand class of hypotheses, and I could just save the idea of a hypothesis until I get there

We consider an additional kind of gap in our probability model. The nature of this gap is: we don't know exactly which order 2 model  $(\mathbb{P}_\square^{X, \mathbb{P}_\square^Y | XD}, A)$  we “ought” to use. To represent this gap we include an unobserved variable  $H$ , the *hypothesis*. We can interpret  $H$  as expressing the fact that, if we knew the value of  $H$  then we would know that our decision problem was represented by a unique order 2 model  $(\mathbb{P}_{h, \square}^{X, \mathbb{P}_{h, \square}^Y | XD}, A)$ . However,  $H$  is not known and in fact we do not know how to determine  $H$  (this is the nature of an *unobserved* variable – there is no process available to find the value it yields). Our model is thus given by

$$(\mathbb{P}_\square^{X|H, \mathbb{P}_\square^Y | HXD}, A)$$

**Definition 3.1** (Ordinary decision problem). An ordinary decision problem  $(\mathbb{P}, \Omega, H, (X, \mathcal{X}), (D, \mathcal{D}), (Y, \mathcal{Y}))$  consists of a fundamental probability set  $\Omega$ , hypotheses  $H : \Omega \rightarrow H$ , observations  $X : \Omega \rightarrow X$ , decisions  $D : \Omega \rightarrow D$  and consequences  $Y : \Omega \rightarrow Y$ , and the latter three random variables are associated with measurement processes. It is equipped with a probability gap model  $\mathbb{P} : \Delta(D)^X \rightarrow \Delta(D)^H$  where  $\Delta(D)^X$  is the set of valid  $D|X$  Markov kernels

$X \rightarrow D$  and  $\Delta(\Omega)^H$  is the set of valid Markov kernels  $H \rightarrow \Omega$ . We require of  $\mathbb{P}$ :

1.  $\mathbb{P}_\alpha^{D|X} = \mathbb{Q}_\alpha^{D|X}$  for all decision functions  $\mathbb{Q}_\alpha^{D|X} \in \Delta(D)^X$
2.  $\mathbb{P}^{X|H} = \mathbb{P}_\alpha^{X|H}$  for all  $\mathbb{P}_\alpha := \mathbb{P}(\mathbb{Q}_\alpha^{D|X})$
3.  $\mathbb{P}^{Y|XDH} = \mathbb{P}_\alpha^{Y|XDH}$  for all  $\mathbb{P}_\alpha := \mathbb{P}(\mathbb{Q}_\alpha^{D|X})$

(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  $\alpha$ .

Under these assumptions  $\mathbb{P}_\square$  is an order 2 model  $(\mathbb{P}_\square^{X, \mathbb{P}_\square^Y|XD}, A)$  which we call a “see-do model”.

I need to update the proof for this claim

### 3.2 Decisions as measurement procedures

We have previously posited that observed variables are variables X – themselves purely mathematical objects – associated with a measurement process  $\mathcal{X}$  that has “one foot in the real world”. In the framework we have proposed here, decisions correspond to a special class of measurement procedure.

Suppose that we are only contemplating decision functions that map deterministically to D. Suppose furthermore that we will D according to a model  $\mathbb{P}_\square$ , a utility function on  $X \times D \times Y \rightarrow \mathbb{R}$  and a decision rule which is a function  $f$  from models, utility functions and decision rules to decisions. Note that models, utility functions and decision rules are all well-defined mathematical objects. If we are confident that our choice will in the end be an element of a well-defined set of objects of the appropriate type, then we are positing that we have a “measurement procedure”  $\mathcal{M}$  that yield models, utilities and decision rules. If so,  $f \circ \mathcal{M}$  – that is, the function that yields a decision – is itself a measurement procedure. This is what is unique about decisions: proposing a complete decision problem with models, utilities and decision rules, defines a measurement procedure for decisions. Other quantities of interest do not seem to have this property – we *require* a measurement process for observations in order to make the whole setup work, but we do not *define* it in the course of setting up a model for our decision problem.



I don't know how important this observation is, but the fact that  $\mathcal{D}$  is an output of a formal decision making system makes it different from other things we might call decisions, and I wonder if I should call it something else in order to avoid ambiguity. The vague reason I think this matters is: whatever you might want to measure, you won't learn more about  $\mathcal{D}$  from it than you already know once you have the model, the utility and the decision rule, this is not a property that other things we call "decisions" share and this distinction might be important regarding judgements of causal contractibility.

### 3.3 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.

Heckerman and Shachter (1995) also explore a decision theoretic approach to causal inference. 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 any notion of conditional independence in our work.

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.

### 3.4 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  $\Theta$ . 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  $R(\theta, \alpha) := \sum_{x \in X} \sum_{d \in D} \mathbb{P}_\theta^X(x) S_\alpha^{D|X}(d|x) l(\theta, d)$ . 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

$$\mathbb{E}_{\mathbb{P}_\alpha}[\mathbf{Y}|h] = \sum_{x \in X} \sum_{d \in D} \sum_{y \in Y} \mathbb{P}^{\mathbf{X}|\mathbf{H}}(x|h) \mathbb{Q}_\alpha^{\mathbf{D}|\mathbf{X}}(d|x) \mathbb{P}^{\mathbf{Y}|\mathbf{H}\mathbf{D}}(y|h, d) \quad (133)$$

$$= \sum_{x \in X} \sum_{d \in D} \mathbb{P}^{\mathbf{X}|\mathbf{H}}(x|h) \mathbb{Q}_\alpha^{\mathbf{D}|\mathbf{X}}(d|x) l(h, d) \quad (134)$$

$$= R(h, \alpha) \quad (135)$$

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

## 4 Repeatable experiments

While there are types of measurement processes we could consider, statistical inference usually proceeds from repeatable measurement processes. A common precise notion of repeatability is the assumption of *exchangeability*. The term “exchangeability”, like the term random variable, is used to refer to assumptions about *measurement processes* as well as properties of *probability models*. If I say a measurement process  $\mathcal{S}$  taking values in  $S^n$  is exchangeable, I might mean:

- I believe that there is some probabilistic model  $(\mathbb{P}, \Omega, \mathcal{F})$  and random variable  $\mathbf{S}$  appropriate for modelling  $\mathcal{S}$  and
  1. The same model is appropriate for any measurement process that first performs  $\mathcal{S}$  and subsequently shuffles the results according to any permutation  $\text{swap}_a : S^n \rightarrow S^n$  or
  2. The same model is appropriate for any measurement process related to  $\mathcal{S}$  by interchanging experimental units or subjects in the real world

On the other hand, if I say a probability model  $(\mathbb{P}, S^{|\mathbf{A}|}, \mathcal{S}^{|\mathbf{A}|})$  is exchangeable, I mean

- For any finite permutation  $\text{swap}_A : S^{|\mathbf{A}|} \rightarrow S^{|\mathbf{A}|}$ ,  $\mathbb{P}^{\mathcal{S} \circ \text{swap}_A} = \mathbb{P}^{\mathcal{S}}$

If I believe a measurement process is exchangeable in the first sense, then this implies that the same probability model is appropriate to model  $\mathcal{S}$  as to model  $\text{swap}_a \circ \mathcal{S}$ , which implies that  $\mathbb{P}^{\mathcal{S}}$  should be an exchangeable probability model. Measurement process exchangeability in the second sense requires us to make explicit the mathematical implications of “interchanging experimental units”, as our semantics of random variables do not say anything about swapping things in the real world. However, the second kind of measurement process exchangeability is more interesting in the context of causal modelling. When we are *acting* on the world, our future actions will often depend on what we have observed in the past, which will often rule out exchangeability in the first

sense. Furthermore, our actions have consequences and so permuting the *labels* associated with actions while not actually changing the actions we take is not a particularly interesting operation. Rather, we are interested in how a model might or might not change if we swap the *actual actions* we take. Swapping experimental units while holding actions constant is one way to achieve this, as it changes the identity of which unit receives which action. See Dawid (2020) and GREENLAND and ROBINS (1986) for further discussions of exchangeability in the context of causal modelling, and note that both authors consider exchanging to be an operation that alters which person receives which treatment.

De Finetti’s well-known representation theorem shows that exchangeable probability models feature a “hypothesis”  $H$  such that the sequence  $S$  is independent and identically distributed conditional on  $H$ . That is: a measurement process that is exchangeable in the first sense should be modelled by a conditionally independent and identically distributed sequence of random variables. The question we want to address here is whether measurement processes that are exchangeable in the second sense imply causal models with particular structure. The answer is yes, although as we discuss the key assumption is *causal contractibility* rather than exchangeability.

In this section, we will consider *do-models*; these are see-do models  $(\mathbb{P}_{\square}^{X|H}, \mathbb{P}_{\square}^{Y|XD^H}, A)$  for which the observations are trivial  $X = *$ . Because observations are things of a different type to consequences – they are not affected by actions – to explore ideas related to symmetries of actions and consequences it is substantially simpler to ignore them. We will investigate how we can add observations to symmetric consequence models. We also assume that the hypotheses are trivial  $H = *$ ; once the decision is chosen, we are left with a single probability model. This also substantially simplifies the arguments to be made.

We will consider two different notions of “repeatable experiments”. Both require a sequence of “decisions” to be made and a sequence of consequences, and we assume that each decision corresponds to a single consequence. One could think about these paired sequences as a series of experiments each with different setting choices available; the decisions are the setting choices and the consequences are the results of each experiment. The first notion we consider will be *commutativity of exchange* – we consider the same model appropriate if we alter our experiment by swapping the experimental settings, or if we make analogous swaps to the experimental results. This assumption could be considered a version of the assumption that experimental units can be interchanged. Consider an experiment involving handing out money or not to person A or person B. Commutativity of exchange says that we should use the same probability model to represent the following two predictions:

- Applying choice 1 to A and choice 2 to B and predicting the vector (consequences for A, consequences for B)
- Applying choice 2 to A and choice 1 to B and predicting the vector (consequences for B, consequences for A)

Under the assumption of commutativity of exchange, consequences of deci-

sions for one “experimental unit” may still depend on decisions made for other “experimental units”. Consider again the experiment above, except instead of two people we are considering giving money to everyone in a particular country. Supposing we don’t otherwise know much about the people we are giving money to, it might be reasonable to posit that a model of the consequences should observe commutativity of exchange. However, giving money to A as well as everyone else will have different consequences for A than giving money to A and no-one else; in the former case, we will create more inflation than in the latter.

The second notion of “repeatable experiments” is *causal contractibility*, a strictly stronger assumption than commutativity of exchange. Causal contractibility is the assumption that, given two different sequences of decisions, the marginal model of consequences corresponding to matching subsequences of decisions will be equal. A causally contractible model says that, if I make the same choice for any subcollection of experiments, I expect the same results from those experiments regardless of whatever choices I make elsewhere.

#### 4.1 Assumptions of repeatability applicable to models of decisions and consequences

In this section we formalise the notion of commutativity of exchange and causal contractibility. We will then go on to prove two representation theorems for causally contractible models – firstly, that they can be represented with a tabular probability model and a lookup function, a construction that is very similar to the kinds of causal models employed by the potential outcomes framework. Secondly, we will show that contractible causal models can also be represented by jointly independent repetitions of a “unit-level consequence map”, indexed by a hypothesis  $H$ .

To begin with, we will define *do* models, which are see-do models with nothing to see.

**Definition 4.1** (Do model). A *do model* is an infinite sequential probability gap model  $(\mathbb{P}_{\square}^{Y||D}, R)$  where  $R$  is a subset of the *ignorant* combs: for all  $n$ ,  $\mathbb{P}_{\alpha}^{D_{[n]}||Y_{[n-1]}} = \text{erase}_{Y_{n-1}} \otimes \mathbb{P}_{\alpha}^{D_{[n]}}$  for some  $\mathbb{P}_{\alpha}^{D_{[n]}} \in \Delta(D^n)$ . That is, we don’t permit any decision  $D_i$  to depend on prior observations  $Y_i$ s.

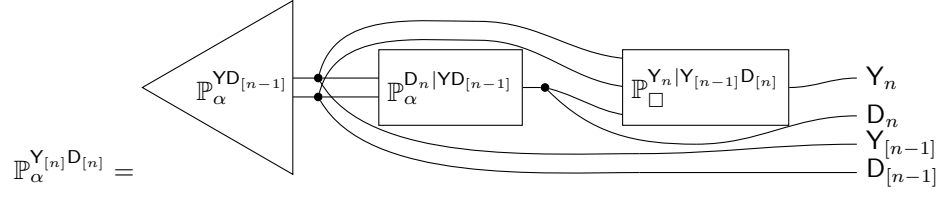
Do models are useful because the  $n$  – *comb*  $\mathbb{P}_{\square}^{Y_{[n]}||D_{[n]}}$  is also the conditional  $\mathbb{P}_{\square}^{Y||D}$ .

**Theorem 4.2** (Existence of marginal). *Given a do model  $(\mathbb{P}_{\square}^{Y||D}, R)$ , for all  $\alpha \in R$ ,  $n \in \mathbb{N}$*

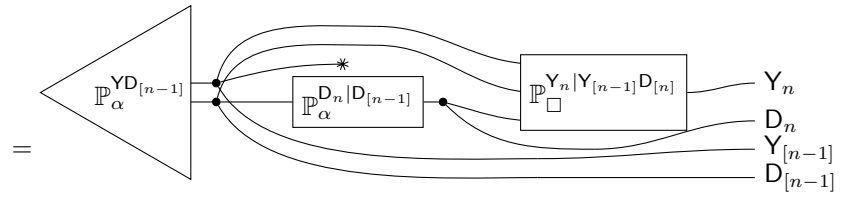
$$\mathbb{P}_{\alpha}^{Y_{[n]}D_i} = \mathbb{P}_{\alpha}^{D_{[n]}} \odot \mathbb{P}_{\square}^{Y_{[n]}||D_{[n]}} \quad (136)$$

*That is,  $\mathbb{P}_{\square}^{Y_{[n]}||D_{[n]}} \cong \mathbb{P}_{\square}^{Y_{[n]}|D_{[n]}}$*

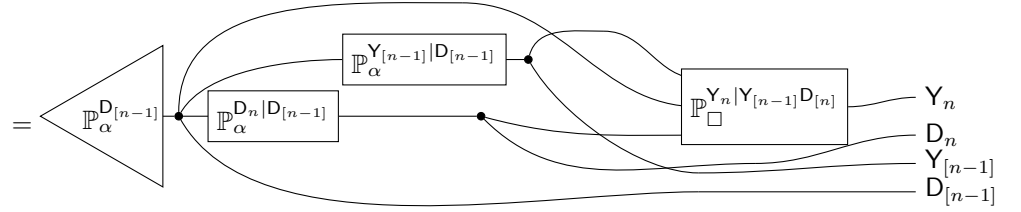
*Proof.* For any  $n > 1 \in \mathbb{N}$ ,  $\alpha \in R$



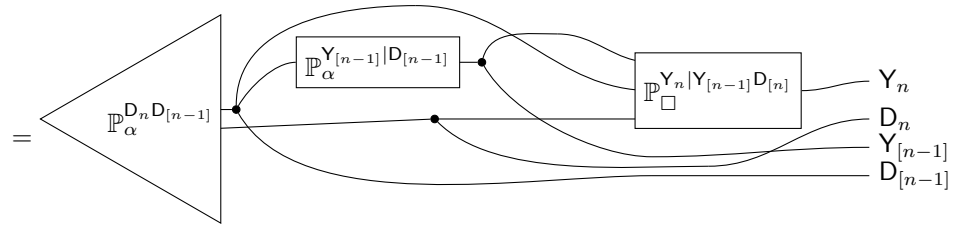
(137)



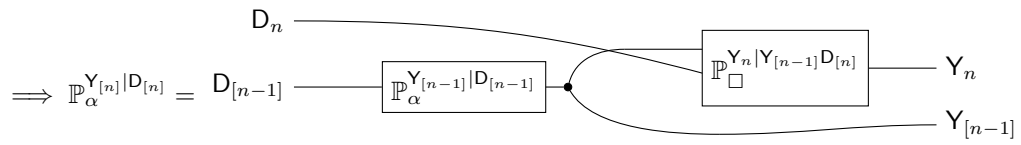
(138)



(139)



(140)



(141)

□

A do model “commutes with exchange” if exchanging decisions or exchanging consequences yields the same model for any finite permutation. The term *commute* comes from the notion that we can apply the exchange before the model  $\mathbb{P}$  or after it and get the same result.

**Definition 4.3** (Commutativity of exchange). Suppose we have a fundamental probability set  $\Omega$  and a do model  $(\mathbb{P}, \mathbf{D}, \mathbf{Y}, R)$  such that  $\mathbf{D} := (\mathbf{D}_i)_{i \in \mathbb{N}}$  and  $\mathbf{Y} := (\mathbf{Y}_i)_{i \in \mathbb{N}}$ . For a finite permutation  $\rho : \mathbb{N} \rightarrow \mathbb{N}$ , define  $\text{swap}_{\rho(D)} : D \rightarrow D$  by  $(d_i)_{i \in \mathbb{N}} \mapsto \delta_{(d_{\rho(i)})_{i \in \mathbb{N}}}$  and  $\text{swap}_{\rho(D \times Y)} : D \times Y \rightarrow D \times Y$  by  $(x_i)_{i \in \mathbb{N}} \mapsto \delta_{(x_{\rho(i)})_{i \in \mathbb{N}}}$ . If, for any two decision rules  $\alpha, \beta \in R$ ,

$$\mathbb{P}_\alpha^{\mathbf{D}} \text{swap}_{\rho(D)} = \mathbb{P}_\beta^{\mathbf{D}} \quad (142)$$

$$\implies \mathbb{P}_\alpha \text{swap}_{\rho(D \times Y)} = \mathbb{P}_\beta \quad (143)$$

Then  $\mathbb{P}$  *commutes with exchanges*.

A do model is causally contractible if it gives identical results for any identical subsequences of two decisions when we limit our attention to the corresponding subsequences of consequences. For example, if we have  $\mathbf{D} = (\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3)$  and  $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3)$  and  $\mathbb{P}_\alpha^{\mathbf{D}_1 \mathbf{D}_3} = \mathbb{P}_\beta^{\mathbf{D}_3 \mathbf{D}_2}$  then  $\mathbb{P}_\alpha^{\mathbf{Y}_1 \mathbf{Y}_3} = \mathbb{P}_\beta^{\mathbf{Y}_3 \mathbf{Y}_2}$ .

**Definition 4.4** (Causal contractibility). Suppose we have a fundamental probability set  $\Omega$  and a do model  $(\mathbb{P}, \mathbf{D}, \mathbf{Y}, R)$  such that  $\mathbf{D} := (\mathbf{D}_i)_{i \in \mathbb{N}}$  and  $\mathbf{Y} := (\mathbf{Y}_i)_{i \in \mathbb{N}}$ . For any  $A = (s_i)_{i \in A}$ ,  $T = (t_i)_{i \in A}$ ,  $A \subset \mathbb{N}$  and  $i < j \implies p_i < p_j$  &  $q_i < q_j$ , let  $\mathbf{D}_S := (\mathbf{D}_i)_{i \in S}$  and  $\mathbf{D}_T := (\mathbf{D}_i)_{i \in T}$ . If for any  $\alpha, \beta \in R$

$$\mathbb{P}_\alpha^{\mathbf{D}_S} = \mathbb{P}_\beta^{\mathbf{D}_T} \implies \mathbb{P}_\alpha^{(\mathbf{D}_i, \mathbf{Y}_i)_{i \in S}} = \mathbb{P}_\beta^{(\mathbf{D}_i, \mathbf{Y}_i)_{i \in T}} \quad (144)$$

then  $\mathbb{P}$  is *causally contractible*.

**Theorem 4.5.** *Causal contractibility implies commutativity of exchange.*

*Proof.* Consider  $\alpha, \beta \in R$  such that  $\mathbb{P}_\alpha^{\mathbf{D}} \text{swap}_{\rho(D)} s = \mathbb{P}_\beta^{\mathbf{D}}$ . Then  $\mathbb{P}_\alpha^{\mathbf{D}_{\rho(\mathbb{N})}} = \mathbb{P}_\beta^{\mathbf{D}}$  also. For any finite  $A \subset \mathbb{N}$

$$\mathbb{P}_\alpha^{\mathbf{D}_{\rho(A)}} = \mathbb{P}_\beta^{\mathbf{D}_A} \quad (145)$$

so by piecewise replicability

$$\mathbb{P}_\alpha^{(\mathbf{D}_i, \mathbf{Y}_i)_{i \in \rho(A)}} = \mathbb{P}_\beta^{(\mathbf{D}_i, \mathbf{Y}_i)_{i \in A}} \quad (146)$$

Thus by Kolmogorov’s extension theorem

$$(\mathbf{D}_i, \mathbf{Y}_i)_{i \in \rho(\mathbb{N})} = \mathbb{P}_\alpha^{(\mathbf{D}_i, \mathbf{Y}_i)_{i \in \mathbb{N}}} \text{swap}_{\rho(D \times A)} \quad (147)$$

$$= \mathbb{P}_\beta^{(\mathbf{D}_i, \mathbf{Y}_i)_{i \in \mathbb{N}}} \quad (148)$$

□

Commutativity of exchange does not imply causal contractibility. For example, suppose  $|D| = 2$ ,  $D = Y = \{0, 1\}$  and we have a do-model  $\mathbb{P}$  such that for all  $\alpha \in R$

$$\mathbb{P}_\alpha^{Y_1 Y_2 | D_1 D_2}(y_1, y_2 | d_1, d_2) = \mathbb{I}((y_1, y_2) = (d_1 + d_2, d_1 + d_2)) \quad (149)$$

Then  $\mathbb{P}_{00}^{Y_1}(y_1) = \mathbb{I}(y_1 = 0)$  while  $\mathbb{P}_{01}^{Y_1} = \mathbb{I}(y_1 = 1)$ , so  $\mathbb{P}$  is not piecewise replicable. However, taking  $(d_i, d_j)$  to be the decision function that deterministically chooses  $(d_i, d_j)$ ,

$$\mathbb{P}_{d_2, d_1}^{Y_1 Y_2 | D_1 D_2}(y_1, y_2) = \mathbb{I}((y_1, y_2) = (d_2 + d_1, d_2 + d_1)) \quad (150)$$

$$= \mathbb{I}((y_2, y_1) = (d_1 + d_2, d_1 + d_2)) \quad (151)$$

$$= \mathbb{P}_{d_1, d_2}^{Y_1 Y_2 | D_1 D_2}(y_2, y_1) \quad (152)$$

so  $\mathbb{P}$  commutes with exchange.

There is a representation theorem for models that commute with exchange which implies that for  $\mathbb{P}$  that commutes with exchange,  $Y_i \perp\!\!\!\perp_{\mathbb{P}} (D_j, Y_j)_{j \in \mathbb{N} \setminus \{i\}} | HD_i$ , where  $H$  is a symmetric function of  $(Y_i, D_i)_{i \in \mathbb{N}}$ .

## 4.2 Representations of contractible probability models

We prove two representation theorems for causally contractible do models. Theorem 4.7 shows that a do model is contractible if and only if it can be represented with a contractible probability distribution over a “table of variables” and a lookup function. This is interesting in its own right, as tabular probability distributions and lookup functions are core elements of the potential outcomes approach. Furthermore, we make use of this theorem in proving Theorem 4.9, which shows a do model is contractible if and only if it can be represented by independent copies of a unit level consequence map jointly parametrised by a hypothesis. We will argue in the next section that jointly parametrised consequence maps are fundamental to all approaches to causal inference.

matrix of variables?

**Definition 4.6** (Contractible probability distribution). Given a fundamental probability set  $\Omega$ , variable  $\mathbf{X} := (\mathbf{X}_i)_{i \in \mathbb{N}}$  and a probability distribution  $\mathbb{P}^{\mathbf{X}} \in \Delta(X^{\mathbb{N}})$ , any  $S = (s_i)_{i \in A}$ ,  $T = (t_i)_{i \in A}$  with  $A \subset \mathbb{N}$  and  $i < j \implies s_i < s_j \wedge t_i < t_j$ , let  $\mathbf{X}_S := (\mathbf{X}_i)_{i \in S}$  and  $\mathbf{X}_T := (\mathbf{X}_i)_{i \in T}$ . If

$$\mathbb{P}^{\mathbf{X}_S} = \mathbb{P}^{\mathbf{X}_T} \quad (153)$$

$\mathbb{P}$  is contractible.

If we have a do model  $\mathbb{P}$  that is causally contractible, we can represent it as an exchangeable probability distribution and a lookup function.

The following can be deduced from the theorems after it, but I thought it might be helpful to have the explanation.



That is, we can define a variable  $\mathbf{Y}^D : \Omega \rightarrow Y^{D \times \mathbb{N}}$  which can be represented as a matrix of variables  $\mathbf{Y}_{ij}$

$$\mathbf{Y}^D = \begin{array}{c} \begin{array}{c} \updownarrow \\ |D| \text{ rows} \end{array} \begin{array}{cccc} \begin{array}{c} \leftarrow \text{N columns} \rightarrow \\ \mathbf{Y}_{11} \quad \mathbf{Y}_{12} \quad \mathbf{Y}_{13} \quad \mathbf{Y}_{14} \\ \mathbf{Y}_{21} \quad \mathbf{Y}_{22} \quad \mathbf{Y}_{23} \quad \mathbf{Y}_{24} \quad \cdots \\ \mathbf{Y}_{31} \quad \mathbf{Y}_{32} \quad \mathbf{Y}_{33} \quad \mathbf{Y}_{34} \end{array} \end{array} \quad (154)$$

and, given any deterministic decision function  $\delta_d$ ,  $d = (d_i)_{i \in \mathbb{N}} \in D^{\mathbb{N}}$ , we can find  $\mathbb{P}^{\mathbf{Y}^{\mathbb{D}}}$  by “looking up”  $d$  in the table. For example, if  $d = (1, 2, 3, 2, \dots)$ , Equation 155 illustrates the idea of “looking up” the relevant elements of  $\mathbf{Y}^D$  and Equation 156 illustrates the resulting value of  $\mathbb{P}^{\mathbf{Y}^{\mathbb{D}}}$ .

$$\begin{array}{ccccccc}
d = & \textcircled{1} & \textcircled{2} & \textcircled{3} & \textcircled{2} & \dots & \\
& \textcircled{Y_{11}} & Y_{12} & Y_{13} & Y_{14} & & \\
Y^D = & Y_{21} & \textcircled{Y_{22}} & Y_{23} & \textcircled{Y_{24}} & \dots & \\
& Y_{31} & Y_{32} & \textcircled{Y_{33}} & Y_{34} & & 
\end{array} \quad (155)$$

$$\mathbb{P}^{\mathbf{Y}|\mathbf{D}}(y|(1, 2, 3, 2, \dots)) = \mathbb{P}^{\mathbf{Y}_1 \mathbf{1} \mathbf{Y}_2 \mathbf{2} \mathbf{Y}_{33} \mathbf{Y}_{24} \dots}(y) \quad (156)$$

The contractibility of  $\mathbb{P}^{\mathbf{Y}^D}$  means that any two subcollections of columns of the same size are equal in distribution, and the exchangeability of  $\mathbb{P}^{\mathbf{Y}^D}$  means that the random variable obtained by permuting its columns is also equal in distribution to  $\mathbf{Y}^D$ .

This representation is very similar to the potential outcomes representation of causal models, with two points of friction. Firstly, we used the assumption of contractibility to derive the contractible table representation, and so we make no claims about what kind of do-model is represented by a non-contractible table lookup. Secondly, we do not yet include any notion of observations, which is a key element of potential outcomes models.

**Theorem 4.7** (Table representation of causally contractible do models). *Suppose we have a fundamental probability set  $\Omega$  and a do model  $(\mathbb{P}, \mathbf{D}, \mathbf{Y}, R)$  such that  $\mathbf{D} := (\mathbf{D}_i)_{i \in \mathbb{N}}$  and  $\mathbf{Y} := (\mathbf{Y}_i)_{i \in \mathbb{N}}$ .  $\mathbb{P}$  is causally contractible if and only if*

$$\mathbb{P}^{Y|D} = \begin{array}{c} \triangle \text{ (containing } \mathbb{P}^{Y^D}\text{)} \\ \text{---} \text{ (from } \mathbb{P}^{Y^D}\text{ to } \mathbb{L}^{D,Y^D}\text{)} \\ \text{---} \text{ (from } D \text{ to } \mathbb{L}^{D,Y^D}\text{)} \\ \square \text{ (containing } \mathbb{L}^{D,Y^D}\text{)} \\ \text{---} \text{ (from } \mathbb{L}^{D,Y^D}\text{ to } Y\text{)} \end{array} \quad (157)$$

$$\Longleftrightarrow \quad (158)$$

$$\mathbb{P}^{Y|D}(y|d) = \mathbb{P}^{(Y_{d_i i}^D)_{\mathbb{N}}}(y) \quad (159)$$

Where  $\mathbb{P}^{\mathbf{Y}^D}$  is a contractible probability measure on  $Y^{D \times \mathbb{N}}$  with respect to the sequence  $\mathbf{Y}^D := (\mathbf{Y}_{ij}^D)_{i \in D, j \in \mathbb{N}}$  and  $\mathbb{L}^{D, \mathbf{Y}^D}$  is the Markov kernel associated with the lookup function

$$l : D^{\mathbb{N}} \times Y^{D \times \mathbb{N}} \rightarrow Y \quad (160)$$

$$((d_i)_{i \in \mathbb{N}}, (y_{ij})_{i \in D, j \in \mathbb{N}}) \mapsto y_{d_i i} \quad (161)$$

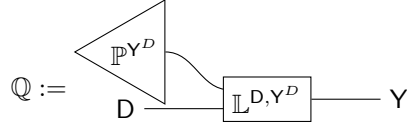
*Proof.* Only if: Choose  $e := (e_i)_{i \in \mathbb{N}}$  such that  $e_{|D|+j}$  is the  $i$ th element of  $D$  for all  $i, j \in \mathbb{N}$ . Abusing notation, write  $e$  also for the decision function that chooses  $e$  deterministically.

Define

$$\mathbb{P}^{\mathbf{Y}^D}((y_{ij})_{D \times \mathbb{N}}) := \mathbb{P}_e^{\mathbf{Y}}((y_{|D|+j})_{i \in D, j \in \mathbb{N}}) \quad (162)$$

Now consider any  $d := (d_i)_{i \in \mathbb{N}} \in D^{\mathbb{N}}$ . By definition of  $e$ ,  $e_{|D|+i} = d_i$  for any  $i, j \in \mathbb{N}$ .

$$\mathbb{Q} : D \rightarrow Y \quad (163)$$



$$\mathbb{Q} := \quad (164)$$

and consider some ordered sequence  $A \subset \mathbb{N}$  and  $B := ((|D|d_i + i))_{i \in A}$ . Note that  $e_B := (e_{|D|d_i + i})_{i \in B} = d_A = (d_i)_{i \in A}$ . Then

$$\sum_{y \in Y^{-1}(y_A)} \mathbb{Q}(y|d) = \sum_{y \in Y^{-1}(y_A)} \mathbb{P}^{(\mathbf{Y}_{d_i}^D)^A}(y) \quad (165)$$

$$= \sum_{y \in Y^{-1}(y_A)} \mathbb{P}_e^{(\mathbf{Y}_{|D|d_i + i})^A}(y) \quad (166)$$

$$= \mathbb{P}_e^{\mathbf{Y}_B}(y_A) \quad (167)$$

$$= \mathbb{P}_d^{\mathbf{Y}_A}(y_A) \quad \text{by causal contractibility} \quad (168)$$

Because this holds for all  $A \subset \mathbb{N}$ , by the Kolmogorov extension theorem

$$\mathbb{Q}(y|d) = \mathbb{P}_d^{\mathbf{Y}}(y) \quad (169)$$

Because  $d$  is the decision function that deterministically chooses  $d$ , for all  $d \in D$

$$\mathbb{Q}(y|d) = \mathbb{P}_d^{\mathbf{Y}^D}(y|d) \quad (170)$$

And because  $\mathbb{P}_d^{\mathbf{Y}|\mathbf{D}}(y|d)$  is unique for all  $d \in D^{\mathbb{N}}$  and  $\mathbb{P}^{\mathbf{Y}|\mathbf{D}}$  exists by assumption

$$\mathbb{P}^{\mathbf{Y}|\mathbf{D}} = \mathbb{Q} \quad (171)$$

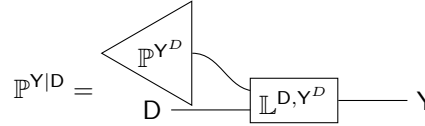
Next we will show  $\mathbb{P}^{\mathbf{Y}^D}$  is contractible. Consider any subsequences  $\mathbf{Y}_S^D$  and  $\mathbf{Y}_T^D$  of  $\mathbf{Y}^D$  with  $|S| = |T|$ . Let  $\rho(S)$  be the “expansion” of the indices  $S$ , i.e.  $\rho(S) = (|D|i + j)_{i \in S, j \in D}$ . Then by construction of  $e$ ,  $e_{\rho(S)} = e_{\rho(T)}$  and therefore

$$\mathbb{P}^{\mathbf{Y}_S^D} = \mathbb{P}_e^{\mathbf{Y}_{\rho(S)}} \quad (172)$$

$$= \mathbb{P}_e^{\mathbf{Y}_{\rho(T)}} \quad \text{by contractibility of } \mathbb{P} \text{ and the equality } e_{\rho(S)} = e_{\rho(T)} \quad (173)$$

$$= \mathbb{P}^{\mathbf{Y}_T^D} \quad (174)$$

If: Suppose



$$\mathbb{P}^{\mathbf{Y}|\mathbf{D}} = \quad (175)$$

and consider any two deterministic decision functions  $d, d' \in D^{\mathbb{N}}$  such that some subsequences are equal  $d_S = d'_S$ .

Let  $\mathbf{Y}^{d_S} = (\mathbf{Y}_{d_i i})_{i \in S}$ .

By definition,

$$\mathbb{P}^{\mathbf{Y}_S|\mathbf{D}}(y_S|d) = \sum_{y_S^D \in Y^{|\mathbf{D}| \times |S|}} \mathbb{P}^{\mathbf{Y}_S^D}(y_S^D) \mathbb{L}^{\mathbf{D}_S, \mathbf{Y}^S}(y_S|d, y_S^D) \quad (176)$$

$$= \sum_{y_S^D \in Y^{|\mathbf{D}| \times |T|}} \mathbb{P}^{\mathbf{Y}_T^D}(y_S^D) \mathbb{L}^{\mathbf{D}_S, \mathbf{Y}^S}(y_S|d, y_S^D) \quad \text{by contractibility of } \mathbb{P}^{\mathbf{Y}_T^D} \quad (177)$$

$$= \mathbb{P}^{\mathbf{Y}_T|\mathbf{D}}(y_S|d) \quad (178)$$

□

Note that in some versions of potential outcomes, for example Rubin (2005), potential outcomes are defined as table-and-lookup models, except without the assumption that the probability distribution over the table is contractible. We speculate that this corresponds to the assumption that if two decisions  $d, d'$  correspond on *the same subset of indices*  $d_S = d'_S$  then  $\mathbb{P}_d^{\mathbf{Y}_S} = \mathbb{P}_{d'}^{\mathbf{Y}_S}$ . This is a weaker assumption than causal contractibility – contractibility requires that we can equate consequence models when any subsequence of the first decision matches any subsequence of the second, while this condition only requires that consequence models are equal when a subsequence of the first decision matches

the same subsequence of the second. It is an open question whether this correspondence does in fact hold.

I think that perhaps the “tidiest” set of assumptions is to consider the above assumption to express the notion of “same local action->same local consequences” and exchangeability to express interchangeability of choices. Contractibility might then be the conjunction of both assumptions.

Theorem 4.8 establishes a claim made earlier: that contractibility is strictly stronger than commutativity of exchange.

**Theorem 4.8.** *Causal contractibility implies commutativity of exchange.*

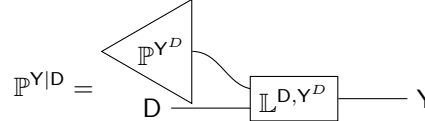
*Proof.* Given a finite permutation  $\rho : \mathbb{N} \rightarrow \mathbb{N}$  and any sequence  $x := (x_i)_{i \in \mathbb{N}}$  let  $\rho(x) = (x_{\rho(i)})_{i \in \mathbb{N}}$  or equivalently  $(x_i)_{i \in \rho(\mathbb{N})}$ . Then for any  $d = (d_i)_{i \in \mathbb{N}}$  and  $y^D := (y_{ij})_{i \in D, j \in \mathbb{N}}$ :

$$l(\rho(d), y^D) = (y_{d_{\rho(i)}})_{i \in \mathbb{N}} \quad (179)$$

$$= (y_{d_i \rho^{-1}(i)})_{i \in \rho(\mathbb{N})} \quad (180)$$

$$= \rho(l(d, \rho^{-1}(y^D))) \quad (181)$$

Suppose we have a fundamental probability set  $\Omega$  and a do model  $(\mathbb{P}, D, Y, R)$  with  $D := (D_i)_{i \in \mathbb{N}}$  and  $Y := (Y_i)_{i \in \mathbb{N}}$  and  $\mathbb{P}$  causally contractible. Then



$$\mathbb{P}^{Y|D} = \begin{array}{c} \triangle \\ \text{P}^{Y^D} \\ \text{D} \longrightarrow \text{L}^{D, Y^D} \longrightarrow Y \end{array} \quad (182)$$

For contractible  $\mathbb{P}^{Y^D}$ . Therefore  $\mathbb{P}^{Y^D}$  is also exchangeable (Kal 2005). But then, given a decision function  $d$  and a finite permutation  $\rho : \mathbb{N} \rightarrow \mathbb{N}$

$$\mathbb{P}_{\rho(d)}^Y(y) = \sum_{y'^D \in Y^{D \times \mathbb{N}}} \mathbb{I}[l_{DY}(\rho(d), y'^D) = y] \mathbb{P}^{Y^D}(y'^D) \quad (183)$$

$$= \sum_{y'^D \in Y^{D \times \mathbb{N}}} \mathbb{I}[l_{DY}(d, \rho^{-1}(y'^D)) = \rho^{-1}(y)] \mathbb{P}^{Y^D}(y'^D) \quad (184)$$

$$= \sum_{y'^D \in Y^{D \times \mathbb{N}}} \mathbb{I}[l_{DY}(d, \rho^{-1}(y'^D)) = \rho^{-1}(y)] \mathbb{P}^{Y^D}(\rho^{-1}(y'^D)) \quad (185)$$

$$= \mathbb{P}_{\rho(d)}^Y(\rho^{-1}(y)) \quad (186)$$

□

We can also represent contractible do-models as a Markov kernels that map from decisions to probability distributions over consequences copied  $\mathbb{N}$  times and jointly parametrised by a hypothesis  $H$ .

**Theorem 4.9.** Suppose we have a fundamental probability set  $\Omega$  and a do model  $(\mathbb{P}, D, Y, R)$  such that  $D := (D_i)_{i \in \mathbb{N}}$  and  $Y := (Y_i)_{i \in \mathbb{N}}$ .  $\mathbb{P}$  is causally contractible if and only if there exists some  $H : \Omega \rightarrow H$  such that  $\mathbb{P}^{Y_i | HD_i}$  exists for all  $i \in \mathbb{N}$  and

$$\mathbb{P}^{Y | HD} = \begin{array}{c} \text{H} \\ \text{D} \end{array} \begin{array}{|c|} \hline \begin{array}{c} \text{H} \\ \text{D} \end{array} \begin{array}{|c|} \hline \text{H}_i \\ \hline \end{array} \begin{array}{|c|} \hline \mathbb{P}^{Y_0 | HD_0} \\ \hline \end{array} \text{Y}_i \\ \hline i \in \mathbb{N} \end{array} \quad (187)$$

$$\iff \quad (188)$$

$$Y_i \perp\!\!\!\perp_{\mathbb{P}} Y_{\mathbb{N} \setminus i}, D_{\mathbb{N} \setminus i} | HD_i \quad \forall i \in \mathbb{N} \quad (189)$$

$$\wedge \mathbb{P}^{Y_i | HD_i} = \mathbb{P}^{Y_0 | HD_0} \quad \forall i \in \mathbb{N} \quad (190)$$

*Proof.* If: By the assumptions of independence and identical conditionals, for any deterministic decision functions  $d, d' \in D$  with equal subsequences  $d_S = d'_T$

$$\mathbb{P}_d^{Y_S | HD}(y|d) = \int_H \prod_{i \in S} \mathbb{P}^{Y_0 | HD_0}(y_i | h, d_i) d\mathbb{P}^H(h) \quad (191)$$

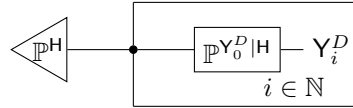
$$= \int_H \prod_{i \in T} \mathbb{P}^{Y_0 | HD_0}(y_i | h, d'_i) d\mathbb{P}^H(h) \quad \text{by equality of subsequences} \quad (192)$$

$$= \mathbb{P}_{d'}^{Y_T | HD}(y|d) \quad (193)$$

Only if: We have

$$\mathbb{P}^{Y | D} = \begin{array}{c} \text{D} \end{array} \begin{array}{|c|} \hline \begin{array}{c} \text{D} \end{array} \begin{array}{|c|} \hline \mathbb{P}^{Y^D} \\ \hline \end{array} \begin{array}{|c|} \hline \mathbb{L}^{D, Y^D} \\ \hline \end{array} \text{Y} \\ \hline \end{array} \quad (194)$$

Also, by contractibility of  $\mathbb{P}^{Y^D}$  and De Finetti's theorem, there is some  $H$  such that



$$\mathbb{P}^{Y^D} = \quad (195)$$

In particular, let  $Y_{\cdot i}^D := (Y_{ji}^D)_{j \in D}$  and  $Y_{\cdot \{i\}^C}^D = (Y_{jk}^D)_{j \in D, k \in \mathbb{N} \setminus \{i\}}$ , and

$$Y_{\cdot i}^D \perp\!\!\!\perp_{\mathbb{P}} Y_{\cdot \{i\}^C}^D | H \quad \text{representation theorem} \quad (196)$$

$$Y^D H \perp\!\!\!\perp_{\mathbb{P}} D \quad \text{by Theorem 2.35 and existence of } \mathbb{P}^{Y^D H} \quad (197)$$

$$Y_{\cdot i}^D \perp\!\!\!\perp_{\mathbb{P}} D | Y_{\cdot \{i\}^C}^D H \quad \text{weak union on Eq. 197} \quad (198)$$

$$Y_{\cdot i}^D \perp\!\!\!\perp_{\mathbb{P}} D Y_{\cdot \{i\}^C}^D | H \quad \text{contraction on Eqs. 196 and 197} \quad (199)$$

$$Y_{\cdot i}^D \perp\!\!\!\perp_{\mathbb{P}} D_{\{i\}^C} Y_{\cdot \{i\}^C}^D | H D_i \quad \text{weak union on Eq. 199} \quad (200)$$

$$D_i \perp\!\!\!\perp_{\mathbb{P}} Y_{\cdot \{i\}^C}^D D_{\{i\}^C} | H D_i Y_{\cdot i}^D \quad \text{due to conditioning on } D_i \quad (201)$$

$$Y_{\cdot i}^D D_i \perp\!\!\!\perp_{\mathbb{P}} D_{\{i\}^C} Y_{\cdot \{i\}^C}^D | H D_i \quad \text{contraction on Eqs. 200 and 201} \quad (202)$$

$$(203)$$

Now, note that  $(Y_i, D_i)$  is a deterministic function of  $(Y_{\cdot i}^D, D_i)$  and  $(Y_{\{i\}^C}, D_{\{i\}^C})$  is a deterministic function of  $(Y_{\{i\}^C}^D, D_{\{i\}^C})$ . Therefore

$$Y_i \perp\!\!\!\perp_{\mathbb{P}} D_{\{i\}^C} Y_{\{i\}^C} | H D_i \quad (204)$$

So, by Theorem 2.35,  $\mathbb{P}^{Y_i | H D_i}$  exists and by contractibility of  $\mathbb{P}^{Y^D}$ , for any  $i, j \in \mathbb{N}$

$$\mathbb{P}^{Y_i | H D_i}(y_i | h, d_i) = \mathbb{P}^{Y_{d_i i}^D | H}(y_i | h) \quad (205)$$

$$= \mathbb{P}^{Y_{d_i j}^D | H}(y_i | h) \quad (206)$$

$$= \mathbb{P}^{Y_j | H D_j}(y_i | h, d_i) \quad (207)$$

□

### 4.3 Extending contractible do models with observations

We have stipulated that, given a do model  $\mathbb{P}_{\square}$  and a decision function  $\alpha$ ,  $\mathbb{P}_{\alpha}$  is a model of the observations given the choice of  $\alpha$ . We can obtain a model of observations and decisions by specifying “half a decision function”. Define an equivalence class  $\sim$  on decision functions  $A$  such that  $\alpha \sim \alpha'$  if  $\mathbb{P}_{\alpha}^{D_1} = \mathbb{P}_{\alpha'}^{D_1}$ . By contractibility,  $\alpha \sim \alpha' \implies \mathbb{P}_{\alpha}^{D_1 Y_1} = \mathbb{P}_{\alpha'}^{D_1 Y_1}$ . Take  $B = A / \sim$ ; then for  $b \in B$  we have  $\mathbb{P}_b^{D_1 Y_1}$  exists.

Fixing some  $b \in B$  fixes a distribution of observations for  $(D_1, Y_1)$  and leaves the rest of the sequence unspecified.

If we are planning to conduct a series of experiments, we think a causally contractible do model is appropriate for these experiments and we have a decision function chosen, then we can restrict the do model to a model of observations by inserting the decision function. If we fix a decision function for the first  $[n]$  experiments and leave the remaining decisions as a probability gap, then we get a see-do model. Combined with assumptions of full support, this see-do model is identifiable. Thus the pair of assumptions

- We can get the same result from performing the same action
- We have a specific plan for which actions to perform

imply identifiability, making them similar in consequence to the assumptions of ignorability (potential outcomes are independent of choices) or causal sufficiency (all causes necessary for identification are observed). However, we may have different reasons for considering these different assumptions to be justified – or not.

There may be many cases where we feel that the first assumption is justifiable but causal identifiability is not licensed. Thus (perhaps surprisingly) the second assumption is quite crucial to the property of identifiability. When we weaken this assumption, we recover familiar rules of causal inference and (non)-identifiability. That is, in this view it is *not* causal contractibility but the existence of a decision function that distinguishes identifiable from non-identifiable models.

Causal identifiability is considered more sound in RCTs, and also people say “surely we can’t define causes by reference to RCTs”. RCTs ensure the existence of a known decision function. We suggest that observational causal inference may proceed from weaker assumptions regarding the decision function.

We can derive the truncated factorisation rule from assuming 1) a causally contractible do model and 2) an unobserved, exchangeable “observational decision function”. This derivation is attractive because

- If we assume there is a unique interventional distribution, it cannot be defined by causal relationships, an observational distribution and the truncated factorisation rule. Why? Because truncated factorisation rule sometimes defines things that don’t exist, and sometimes multiple things satisfy the requirements it imposes
- On the other hand, by the representation theorem above, a unique “IID” interventional distribution  $\iff$  a causally contractible do-model
- Furthermore, causally contractible do-model + exchangeable observational decision rule  $\implies$  truncated factorisation (but not the other way around!)

todo, below is just copied and pasted for now

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  $X \longrightarrow Y$  and  $X \longleftarrow Z \longrightarrow Y$ . “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  $\mathbb{P}_x^{\text{XYZ}}$  on  $X \times Y \times Z$  given by (Pearl, 2009, page 24):

$$\mathbb{P}_x^{\text{XYZ}}(x', y, z) = \bar{\mathbb{P}}^{\text{Y|XZ}}(y|x, z) \mathbb{P}^{\text{Z}}(z) \llbracket x = x' \rrbracket \quad (208)$$

Equation 208 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;  $\mathbb{P}^{\text{Z}} = \mathbb{P}_x^{\text{Z}}$ . Finally, there must be some version of the interventional conditional probability  $\text{Y}|\text{(X, Z)}$  that is equal to some version of the observational conditional probability  $\text{Y}|\text{(X, Z)}$ ; there exists  $\bar{\mathbb{P}}^{\text{Y|XZ}}$  and  $\bar{\mathbb{P}}_x^{\text{Y|XZ}}$  such that  $\bar{\mathbb{P}}^{\text{Y|XZ}} = \bar{\mathbb{P}}_x^{\text{Y|XZ}}$ . We use the overbars here to indicate that, unlike in familiar cases, the particular choice of  $\bar{\mathbb{P}}^{\text{Y|XZ}}$  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  $x \mapsto \mathbb{P}_x^{\text{XYZ}}$  such a model, and if we include random choices we can consider  $\mathbb{Q} : \mathbb{P}_\alpha^{\text{X}} \mapsto \sum_x \mathbb{P}_\alpha^{\text{X}}(x) \mathbb{P}_x^{\text{XYZ}}$  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  $\bar{\mathbb{P}}^{\text{Y|XZ}}$  may be chosen arbitrarily on a set of measure zero with regard to  $\mathbb{P}^{\text{XZ}}$ . As a result, depending on the choice of  $\bar{\mathbb{P}}^{\text{Y|XZ}}$ , Equation 208 can be satisfied by multiple probability distributions that differ in meaningful ways. For example, suppose  $X$ ,  $Y$  and  $Z$  are binary and  $\mathbb{P}((X, Z) \propto (1, 1)) = 1$ . Then we can consistently choose  $\bar{\mathbb{P}}^{\text{Y|XZ}}(1|0, 1) = 1$  or  $\bar{\mathbb{P}}^{\text{Y|XZ}}(1|0, 1) = 0$  because  $\{0, 1\}$  is a measure zero event. However, the first choice gives us  $\mathbb{P}_0^{\text{XYZ}}(0, 1, 1) = 1$  while the second gives us  $\mathbb{P}_0^{\text{XYZ}}(0, 1, 1) = 0$ , 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 208. For example, suppose  $X = f \circ Z$  for some  $f$ . Then we must have  $\mathbb{P}_x^{\text{X}}(x') = \mathbb{P}_x^{\text{Z}}(f^{-1}(x'))$  for any  $x$ . However, we also have  $\mathbb{P}_x^{\text{X}}(x') = \llbracket x = x' \rrbracket$  for all  $x, x'$  and  $\mathbb{P}_x^{\text{Z}} = \mathbb{P}^{\text{Z}}$  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 208 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 interven-



tions and it does not share these problems of non-uniqueness and non-existence with models defined via truncated factorisation.

In our original look at truncated factorisation, we noted a few problems with Equation 208 being a *definition* of interventional probability models. In particular:

- There may be multiple different probability models that satisfy Equation 208 for different versions of the disintegration  $\mathbb{P}^{Y|XZ}$
- There may be no probability models that satisfy Equation 208

We propose a different way to define interventional probability models:

- The interventional probability model is some probability 2-comb

$$\mathbb{Q}^{Z \square Y | X} \quad (209)$$

- For some observational conditional probability  $\mathbb{Q}_{\text{obs}}^{X|Z}$ , observations are distributed according to

$$\mathbb{Q}_{\text{obs}} := \text{insert}(\mathbb{Q}_{\text{obs}}^{X|Z}, \mathbb{Q}^{Z \square Y | X}) \quad (210)$$

Note that, by definition,  $\mathbb{Q}^{Y|XZ}$  exists and all versions of it are also versions of  $\mathbb{Q}_{\text{obs}}^{Y|XZ}$ . If in addition, for every  $x \in X$ , 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 there exists a version of  $\mathbb{Q}_{\text{obs}}^{Y|XZ}$  such that:

$$\mathbb{Q}_x^{XYZ} = \mathbb{Q}_{\text{obs}}^{Y|XZ}(y|x, z) \mathbb{Q}_{\text{obs}}^Z(z) \llbracket x = x' \rrbracket \quad (211)$$

This is similar to Equation 208. The two key differences are that this is existentially quantified over  $\mathbb{Q}_{\text{obs}}^{Y|XZ}$  and we have made explicit the assumptions that hard interventions on  $X$  are valid inserts.

## 5 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.

## 5.1 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 Peters and Bühlmann (2015) for a detailed treatment of when two graphs may or may not imply the same underlying model).

We will adopt the definition discussed in Section ??: we take the causal Bayesian network in question to express the assumption that the result of intervention is modeled by some probability 2-comb  $\mathbb{P}^{W \square Y | X}$  and the observations are distributed according to  $\text{insert}(\mathbb{P}_{\text{obs}}^{X|Z}, \mathbb{P}^{W \square Y | X})$  for some  $\mathbb{P}_{\text{obs}}^{X|Z}$ .

We are uncertain about the particular 2-comb  $\mathbb{P}^{W \square Y | X}$  that we should use to model interventions, as well as the particular  $\mathbb{P}_{\text{obs}}^{X|Z}$  appropriate for observations, so we will represent this uncertainty with an unobserved variable  $H$ . Furthermore, if we are being precise about what is being modeled, we suppose that 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 a fundamental probability set  $\Omega$  (assume  $n < m$ ).

With this additional detail, we interpret Equation 210 as saying

$$\mathbb{Q}_{\text{obs}}^{W_i X_i Y_i | H} = \text{insert}(\mathbb{Q}_{\text{obs}}^{X_j | H W_j}, \mathbb{Q}^{W_j | H \square Y_j | X_j}) \quad (212)$$

for all  $i \in [n]$ ,  $j \in (n, m]$ .

Because we are now considering sequences of observations and consequences, we also think it is reasonable to understand a CBN model as coming equipped with assumptions of mutual independence:

$$V_i \perp\!\!\!\perp_{\mathbb{Q}} V_{[m] \setminus \{i\}} | H \forall i \in [n] \quad (213)$$

$$W_i \perp\!\!\!\perp_{\mathbb{Q}} V_{[m] \setminus \{i\}} | H \forall i \in [m] \quad (214)$$

$$Y_i \perp\!\!\!\perp_{\mathbb{Q}} V_{[m] \setminus \{i\}} | (H, X_i, W_i) \forall i \in [m] \quad (215)$$

$$(216)$$

The first condition says that the observations  $V_{[n]}$  are mutually independent, the second that if we ignore the  $X_i$ s and  $Y_i$ s, then the  $W_i$ s are mutually independent for all  $[m]$  and the third says that the  $Y_i$ s are independent of the other variables in the sequence conditional on  $(H, X_i, W_i)$ . Note that we exclude  $X_i$  from these conditional independence assumptions. The reason for this is that we interpret  $X_i$  as a directly controlled variable and as such it may be chosen to be dependent on other variables in the sequence.

**Definition 5.1** (Order 2 model associated with the CBN in question). A CBN order 2 model  $\mathbb{Q}^{V_{[n]}W_{(n,m)}|H \square Y_{(n,m)}|X_{(n,m)}}$  where  $V_i = (W_i, X_i, Y_i)$  for  $i \in [m]$ , such that the CBN mutual independences hold:

$$V_i \perp\!\!\!\perp_{\mathbb{Q}} V_{[m] \setminus \{i\}} | H \forall i \in [n] \quad (217)$$

$$W_i \perp\!\!\!\perp_{\mathbb{Q}} V_{[m] \setminus \{i\}} | H \forall i \in [m] \quad (218)$$

$$Y_i \perp\!\!\!\perp_{\mathbb{Q}} V_{[m] \setminus \{i\}} | (H, X_i, W_i) \forall i \in [m] \quad (219)$$

$$(220)$$

Under these assumptions  $\mathbb{Q}^{W_i|H}$  and  $\mathbb{Q}^{Y_i|HX_iW_i}$  exist, and for all  $i \in [n]$ ,  $j, k \in [m]$

$$\mathbb{Q}^{W_j|H \square Y_j|X_j} = \mathbb{Q}^{W_k|H \square Y_k|X_k} \quad (221)$$

$$\mathbb{Q}^{W_iX_iY_i|H} = \text{insert}(\mathbb{Q}_{\text{obs}}^{X_j|HW_j}, \mathbb{Q}^{W_j|H \square Y_j|X_j}) \quad (222)$$

for all  $i \in [n]$ ,  $j \in (n, m]$ .

**Theorem 5.2** (Existence of CBN 2-comb). *Given a CBN order 2 model  $\mathbb{Q}^{V_{[n]}W_{(n,m)}|H \square Y_{(n,m)}|X_{(n,m)}}$  in accordance with Definition 5.1, there exist conditional probabilities  $\mathbb{Q}^{W_i|H}$  and  $\mathbb{Q}^{Y_i|HX_iW_i}$  for all  $i$ .*

*Proof.* Order 2 models  $\mathbb{Q}^{V_{[n]}W_{(n,m)}|H \square Y_{(n,m)}|X_{(n,m)}}$  have the following conditional probabilities:

$$\mathbb{Q}^{V_{[n]}W_{(n,m)}|H} \quad \text{exists} \quad (223)$$

$$\mathbb{Q}^{Y_{(n,m)}|X_{(n,m)}W_{(n,m)}V_{[n]}H} \quad \text{exists} \quad (224)$$

Equations 223 and 224 together with conditional independences 217, 218, 219 and Theorem 2.35 imply there exist versions of the following conditional probabilities such that

$$\mathbb{Q}^{\mathbf{V}_i|\mathbf{H}\mathbf{V}_{[m]\setminus\{i\}}} = \mathbb{Q}^{\mathbf{V}_i|\mathbf{H}} \otimes \text{erase}_{V_{m-1}} \quad \forall i \in [n] \quad (225)$$

$$\mathbb{Q}^{\mathbf{W}_i|\mathbf{H}\mathbf{V}_{[m]\setminus\{i\}}} = \mathbb{Q}^{\mathbf{W}_i|\mathbf{H}} \otimes \text{erase}_{V_{m-1}} \quad \forall i \in [m] \quad (226)$$

$$\mathbb{Q}^{\mathbf{Y}_i|\mathbf{H}\mathbf{X}_i\mathbf{W}_i\mathbf{V}_{[m]\setminus\{i\}}} = \mathbb{Q}^{\mathbf{Y}_i|\mathbf{H}\mathbf{X}_i\mathbf{W}_i} \otimes \text{erase}_{V_{m-1}} \quad \forall i \in [m] \quad (227)$$

□

Note also that Equation 221 implies that the conditional probabilities from Theorem 5.2 can be chosen to be equal for all  $i, j \in [m]$ :

$$\mathbb{Q}^{\mathbf{W}_i|\mathbf{H}} = \mathbb{Q}^{\mathbf{W}_j|\mathbf{H}} \quad (228)$$

$$\mathbb{Q}^{\mathbf{Y}_i|\mathbf{X}_i\mathbf{W}_i\mathbf{H}} = \mathbb{Q}^{\mathbf{Y}_j|\mathbf{X}_j\mathbf{W}_j\mathbf{H}} \quad (229)$$

**Theorem 5.3** (CBN observations are identically distributed). *Given a CBN order 2 model  $\mathbb{Q}^{\mathbf{V}_{[n]}\mathbf{W}_{(n,m)}|\mathbf{H}\square\mathbf{Y}_{(n,m)}|\mathbf{X}_{(n,m)}}$  in accordance with Definition 5.1, there exists  $\mathbb{Q}^{\mathbf{V}_i|\mathbf{H}}$  for  $i \in [n]$  and for all  $i, j \in [n]$*

$$\mathbb{Q}^{\mathbf{V}_i|\mathbf{H}} = \mathbb{Q}^{\mathbf{V}_j|\mathbf{H}} \quad (230)$$

*Proof.* Existence follows from the fact that  $\mathbf{V}_i = \pi_i \circ \mathbf{V}_{[n]}$  with  $\pi_i$  the function  $V^n \rightarrow V$  that projects the  $i$ th index. Theorem 2.15 then implies  $\mathbb{Q}^{\mathbf{V}_i|\mathbf{H}} = \mathbb{Q}^{\mathbf{V}_{[n]}|\mathbf{H}}\mathbb{F}_{\pi_i}$ .

Equality follows from the fact that for all  $i, j \in [n]$

$$\mathbb{Q}^{\mathbf{V}_i|\mathbf{H}} = \text{insert}(\mathbb{Q}_{\text{obs}}^{\mathbf{X}_j|\mathbf{H}\mathbf{W}_j}, \mathbb{Q}^{\mathbf{W}_j|\mathbf{H}\square\mathbf{Y}_j|\mathbf{X}_j}) \quad (231)$$

$$= \mathbb{Q}^{\mathbf{V}_j|\mathbf{H}} \quad (232)$$

□

## 5.2 See-do models corresponding to causal Bayesian networks

We have defined a class of order 2 probability gap models associated with causal Bayesian networks. We next want to ask: when does a see-do model induce an order 2 model associated with a CBN? Specifically, given a see-do model  $\mathbb{T}^{\mathbf{V}_{[n]}|\mathbf{H}\square\mathbf{V}_{(n,m)}|\mathbf{D}}$  with decision rules of type  $\{\mathbb{T}_{\alpha}^{\mathbf{D}|\mathbf{V}_{[n]}}\}_{\alpha \in A}$ , when is there a CBN model  $\mathbb{Q}^{\mathbf{V}_{[n]}\mathbf{W}_{(n,m)}|\mathbf{H}\square\mathbf{Y}_{(n,m)}|\mathbf{X}_{(n,m)}}$  with inserts of type  $\{\mathbb{Q}_{\alpha}^{\mathbf{X}_{(n,m)}|\mathbf{V}_{[n]}\mathbf{W}_{(n,m)}\mathbf{H}}\}_{\alpha \in A}$  such that marginalising  $\mathbb{T}_{\alpha}$  over  $\mathbf{D}$  yields  $\mathbb{Q}_{\alpha}$  for all  $\alpha \in A$ ?

$$\mathbb{T}_\alpha^{V_{[m]}|H} = H \rightarrow \boxed{\mathbb{T}^{V_{[n]}|H}} \rightarrow \boxed{\mathbb{T}_\alpha^{D|V_{[n]}}} \rightarrow \boxed{\mathbb{T}^{V_{(n,m)}|DV_A H}} \rightarrow \begin{matrix} V_{[n]} \\ X_{(n,m)} \\ Y_{(n,m)} \\ W_{(n,m)} \end{matrix} \quad (233)$$

$$= H \rightarrow \boxed{Q^{V_{[n]}W_{(n,m)}|H}} \rightarrow \boxed{Q_\alpha^{X_{(n,m)}|HV_{[n]}W_{(n,m)}}} \rightarrow \boxed{Q^{Y_{(n,m)}|X_{(n,m)}W_{(n,m)}H}} \rightarrow \begin{matrix} V_{[n]} \\ X_{(n,m)} \\ Y_{(n,m)} \\ W_{(n,m)} \end{matrix} \quad (234)$$

$$= Q_\alpha^{V_{[m]}|H} \quad (235)$$

We need a number of conditions to hold for this to be true of any  $\mathbb{T}$ ; first,  $\mathbb{T}$  needs to satisfy the CBN versions of “mutually independent” (Eqs. 217-219) as well as the CBN version of “identically distributed” (Eqs. 228, 229 and 230). Finally, we require decisions  $D$  to act as “interventions” on  $X$  in an appropriate way.

Theorem 5.4 shows that this correspondence holds exactly when:

- The CBN mutual independences (Definition 5.1) hold for the  $\mathbb{T}$
- For all  $i \in (n, m]$ ,  $W_i \perp_{\mathbb{T}} D|H$
- For all  $i \in [m]$ ,  $Y_i \perp_{\mathbb{T}} D|(W_i, X_i, H)$ ; we say that under conditions of perfect information,  $(W_i, X_i)$  control  $Y_i$  by proxy

**Theorem 5.4.** *Given a see-do model  $\mathbb{T}^{V_{[n]}|H \square V_{(n,m)}|D}$  there exists a corresponding CBN probability 2-comb  $Q^{V_{[n]}W_{(n,m)}|H \square Y_{(n,m)}|X_{(n,m)}}$  if and only if (1) the CBN mutual independences hold*

$$V_i \perp_{\mathbb{T}} V_{[m] \setminus \{i\}} | H \forall i \in [n] \quad (236)$$

$$W_i \perp_{\mathbb{T}} V_{[m] \setminus \{i\}} | H \forall i \in [m] \quad (237)$$

$$Y_i \perp_{\mathbb{T}} V_{[m] \setminus \{i\}} | H X_i \forall i \in [m] \quad (238)$$

$$(239)$$

(2) for every insert  $\alpha$  the following conditional probabilities are identical

$$\mathbb{T}_\alpha^{V_i|H} = \mathbb{T}_\alpha^{V_j|H} \quad \forall i, j \in [n] \quad (240)$$

$$\mathbb{T}_\alpha^{W_i|H} = \mathbb{T}_\alpha^{W_j|H} \quad \forall i, j \in [m] \quad (241)$$

$$\mathbb{T}_\alpha^{Y_i|X_i W_i H} = \mathbb{T}_\alpha^{Y_j|X_j W_j H} \quad \forall i, j \in [m] \quad (242)$$

(3)  $W_i$  is unaffected by  $D$

$$W_i \perp_{\mathbb{D}} \quad (243)$$

And (4) under conditions of perfect information,  $(W_i, X_i, V_{[n]})$  control  $Y_i$  by proxy:

$$Y_i \perp\!\!\!\perp_{\mathbb{T}} D | (W_i, X_i, H, V_{[n]}) \forall i \in [m] \quad (244)$$

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

$$\mathbb{T}^{V_{[n]} V_j | H D} = \quad (245)$$

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

$$\mathbb{R}_\alpha^{X_j | V_{[n]} W_j H} := \quad (246)$$

Then

$$\quad (247)$$

$$\quad (248)$$

$$\quad (249)$$

**Only if:** Suppose the CBN mutual independences do not hold for  $\mathbb{T}$ . Then there must be some  $\alpha$  such that one of these conditional independences does not hold for  $\mathbb{T}_\alpha$ . By construction of CBN order 2 models, these independences hold for every probability model in the range of every CBN order 2 model  $\mathbb{Q}$ . Thus there is no CBN model corresponding to  $\mathbb{T}$ .

Suppose for some  $\alpha$  and  $i, j \in [n]$  we have  $\mathbb{T}_\alpha^{V_i | H} \neq \mathbb{T}_\alpha^{V_j | H}$ .

Suppose the assumption of proxy control does not hold for  $\mathbb{T}$ . Then there is some  $d, d' \in D$ ,  $w \in W$ ,  $h \in H$ ,  $v \in V^n$ ,  $x \in X$  and  $y \in Y$  such that

$$\mathbb{T}^{Y_j | W_j V_{[n]} H X_j D}(y|w, v, h, x, d) \neq \mathbb{T}^{Y_j | W_j V_{[n]} H X_j D}(y|w, v, h, x, d') \quad (250)$$

$$\text{and } \mathbb{T}_\alpha^{X_j W_j V_{[n]} | H D}(x, w, v|h, d) > 0 \quad (251)$$

$$\text{and } \mathbb{T}_\alpha^{X_j W_j V_{[n]} | H D}(x, w, v|h, d') > 0 \quad (252)$$

$$(253)$$

If Equation 250 only held on sets of measure 0 then we could choose versions of the conditional probabilities such that the independence held.

Then

$$\mathbb{T}_d^{Y_j | W_j V_{[n]} H X_j}(y|w, v, h, x) = \mathbb{T}^{Y_j | W_j V_{[n]} H X_j D}(y|w, v, h, x, d) \quad (254)$$

$$\neq \mathbb{T}^{Y_j | W_j V_{[n]} H X_j D}(y|w, v, h, x, d') \quad (255)$$

$$= \mathbb{T}_{d'}^{Y_j | W_j V_{[n]} H X_j}(y|w, v, h, x) \quad (256)$$

$$\implies \mathbb{T}_d^{Y_j | W_j V_{[n]} H X_j D}(y|w, v, h, x) \neq \mathbb{Q}_d^{Y_j | W_j V_{[n]} H X_j D}(y|w, v, h, x) \quad (257)$$

$$\text{or } \mathbb{T}_{d'}^{Y_j | W_j V_{[n]} H X_j D}(y|w, v, h, x) \neq \mathbb{Q}_{d'}^{Y_j | W_j V_{[n]} H X_j D}(y|w, v, h, x) \quad (258)$$

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  $y \in Y$

$$\mathbb{P}_d^{Y_j | W_j V_{[n]} H X_j D}(y|w, v, h, x) = \mathbb{T}^{Y_j | W_j V_{[n]} H X_j}(y|w, v, h, x) \quad (259)$$

$$\neq \mathbb{U}^{Y_j | W_j V_{[n]} H X_j}(y|w, v, h, x) \quad (260)$$

$$\neq \text{model } Q_d^{Y_j | W_j V_{[n]} H X_j D}(y|w, v, h, x) \quad (261)$$

□

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

### 5.3 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

$$\begin{aligned}
T^{V_{[n]}V_{(n,m)}|HD} &= \begin{array}{c} \begin{array}{c} H \\ D \end{array} \begin{array}{c} \boxed{U^{W_jV_A|H}} \\ \boxed{T^{X|W_jV_AHD}} \end{array} \begin{array}{c} \begin{array}{c} \text{---} \end{array} \\ \begin{array}{c} \text{---} \end{array} \end{array} \begin{array}{c} \begin{array}{c} V_A \\ W_j \end{array} \\ \begin{array}{c} \boxed{T^{Y_j|W_jV_AX_j}} \\ \begin{array}{c} Y_j \\ X_j \end{array} \end{array} \end{array} \quad (262) \\
= \begin{array}{c} \begin{array}{c} H \\ D \end{array} \begin{array}{c} \boxed{U^{W_jV_A|H}} \\ \boxed{T^{X|W_jV_AHD}} \end{array} \begin{array}{c} \begin{array}{c} \text{---} \end{array} \\ \begin{array}{c} \text{---} \end{array} \end{array} \begin{array}{c} \boxed{K} \\ \begin{array}{c} V_A \\ W_j \\ Y_j \\ X_j \end{array} \end{array} = T^{V_{[n]}W_jX_j|HD}\mathbb{M} \quad (263)
\end{aligned}$$

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.

## 6 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  $\mathbf{X}$  taking values in  $X$  are a variable  $\mathbf{Y}^X$  taking values in  $Y^X$  such that  $\mathbf{Y}$  is related to  $\mathbf{Y}^X$  and  $\mathbf{X}$  via a *selector*.

**Definition 6.1** (Selector). Given variables  $\mathbf{X} : \Omega \rightarrow X$  and  $\{\mathbf{Y}^x : \Omega \rightarrow Y \mid x \in X\}$ , define  $\mathbf{Y}^X : (\mathbf{Y}^x)_{x \in X}$ . The selector  $\pi : X \times Y^X \rightarrow Y$  is the function that sends  $(x, y^1, \dots, y^{|X|}) \rightarrow y^x$ .

**Definition 6.2** (Potential outcomes: formal requirement). Given variables  $\mathbf{Y} : \Omega \rightarrow Y$  and  $\mathbf{X} : \Omega \rightarrow X$ , we introduce a collection of latent variables called *potential outcomes*  $\mathbf{Y}^X := (\mathbf{Y}^x)_{x \in X}$  such that  $\mathbf{Y} = \pi \circ (\mathbf{X}, \mathbf{Y}^X)$ . A *potential outcomes model* is any consistent model of  $\mathbf{Y}$ ,  $\mathbf{X}$  and  $\mathbf{Y}^X$ .

Lemma 6.3 shows we can always define trivial potential outcomes of  $\mathbf{Y}$  with respect to  $\mathbf{X}$  by taking the product of  $|X|$  copies of  $\mathbf{Y}$ . We need some other constraint on the values of potential outcomes besides the formal definition 6.2 if we want them to be informative.

**Lemma 6.3** (Trivial formal potential outcomes). *For any variables  $\mathbf{Y} : \Omega \rightarrow Y$ ,  $\mathbf{X} : \Omega \rightarrow X$  and  $\mathbf{W} : \Omega \rightarrow W$ , we can always define potential outcomes  $\mathbf{Y}_X$  such that any consistent model  $\mathbb{K}^{\mathbf{Y}\mathbf{X}|\mathbf{W}}$  can be extended to a consistent model of  $\mathbb{K}^{\mathbf{Y}\mathbf{X}\mathbf{Y}^X|\mathbf{W}}$ .*

*Proof.* Define  $\mathbf{Y}^X := (\mathbf{Y})_{x \in X}$ . Then we can consistently extend  $\mathbb{K}^{\mathbf{Y}\mathbf{X}|\mathbf{W}}$  to  $\mathbb{K}^{\mathbf{Y}\mathbf{X}\mathbf{Y}^X|\mathbf{W}}$  by repeated application of Lemma 2.38.  $\square$

The trivial potential outcomes of Lemma 6.3 are in many cases unsatisfactory for what we want potential outcomes to represent. Thus Definition 6.2 is incomplete. In common with observable variables, the definition of potential outcomes involves both the formal requirement of Definition 6.2, 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  $\mathbf{Y}^x := \pi(x, \mathbf{Y})$  is “the value that  $\mathbf{Y}$  would take if  $\mathbf{X}$  were  $x$ , whether or not  $\mathbf{X}$  actually takes the value  $x$ ”. We will call this additional element of the definition of potential outcomes the *counterfactual extension*.

**Definition 6.4** (What potential outcomes model: counterfactual extension). Given observables  $\mathbf{X}$ ,  $\mathbf{Y}$  and  $\mathbf{Y}^X$ ,  $\mathbf{Y}^X$  are potential outcomes if they satisfy Definition 6.2 and for all  $x \in X$ , the individual potential outcome  $\mathbf{Y}^x := \pi(x, \mathbf{Y})$  models the value  $\mathbf{Y}$  would take if  $\mathbf{X}$  took the value  $x$ .

Because observables resolve to a single canonical value, the conditional in Definition 6.4 is eventually satisfied for exactly one  $x \in X$ , at which point  $\mathbf{Y}^{x'}$  for all  $x' \neq x$  are guaranteed not to resolve. Nevertheless, we can maybe draw some conclusions about  $\mathbf{Y}^X$  from Definition 6.4. For example, it seems unreasonable in light of this definition to assert that  $\mathbf{Y}^x$  is *necessarily* identical to  $\mathbf{Y}$  for all  $x \in X$ , which rules out the strictly trivial potential outcomes of Lemma 6.3.

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 6.2 and 6.4 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 6.4 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

$$\mathbb{P}^{XY^X|H} = H \quad (264)$$

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 264 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 6.5.

**Lemma 6.5** (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*

$$\mathbb{Q} := \begin{array}{c} Y^X \\ X \\ W \end{array} \begin{array}{c} \diagup \\ \diagdown \\ \longrightarrow \end{array} \Pi \longrightarrow Y \quad (265)$$

$$\iff \quad (266)$$

$$\mathbb{Q}_{(y^x, x, w)}^y = \Pi_{(x, y^x)}^y \quad \forall y, y^x, x, w \quad (267)$$

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

*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  $\mathbb{T}^{Y^X X|W}(y, y^X, x|w) = 0$  and so  $\mathbb{T}^{Y|Y^X X W}$  is arbitrary on this collection of values. Equations 265 and 267 are equivalent to the statement  $Y \stackrel{a.s.}{=} \pi \circ (X, Y^X)$ .  $\square$

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

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

**Definition 6.6.** 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 6.2 and is a prechoice variable; that is,  $(Y^X, W) \perp\!\!\!\perp_{\mathbb{T}} D|H$ .

Owing to the subtlety of interpreting Definition 6.4, we don’t know a straightforward argument to the effect that Definition 6.6 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 6.6 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 6.7.** 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  $y$

$$\mathbb{T}^{Y|HWD}(y|h, w, 0, 0) \neq \mathbb{T}^{Y|HWD}(y|h, w, 0, 2) \quad (268)$$

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,

$$\mathbb{P}_0^{Y|HD}(y|h, 0) = \sum_{x \in \{0,1\}, w \in W} \mathbb{T}^{W|H}(w|h) \mathbb{S}_0^{D|W}(0|w) \mathbb{T}^{X|HWD}(x|h, w, 0) \mathbb{T}^{Y|HWD}(y|h, w, x, 0) \quad (269)$$

$$= \mathbb{T}^{Y|HWD}(y|h, w, 0, 0) \quad (270)$$

$$\neq \mathbb{P}_2^{Y|HD} \quad (271)$$

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

$$\mathbb{P}_0^{Y|HD}(y|h, 0) = \sum_{y^X \in Y^2, x \in \{0,1\}} \mathbb{T}^{Y^X|H}(y^X|h) \mathbb{T}^{X|HDY^X}(x|h, 0, y^X) \Pi(y|x, y^X) \quad (272)$$

$$= \sum_{y^X} \mathbb{T}^{Y^X|H}(y^X|h) \Pi(y|0, y^X) \quad (273)$$

$$= \sum_{y^X \in Y^2, x \in \{0,1\}} \mathbb{T}^{Y^X|H}(y^X|h) \mathbb{T}^{X|HDY^X}(x|h, 2, y^X) \Pi(y|x, y^X) \quad (274)$$

$$= \mathbb{P}_2^{Y|HD} \quad (275)$$

Here we use the property  $Y^X \perp\!\!\!\perp_{\mathbb{T}} D|H$ , implied by the assumption that  $Y^X$  is a prechoice variable. Equations 271 and 275 are clearly contradictory, thus there can be no potential outcomes  $Y^X$  in this example.

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

**Theorem 6.8** (Existence of potential outcomes). *Suppose we have a standard decision problem  $\{\mathbb{T}^{WZ|HD}, \{\mathbb{S}_\alpha\}_{\alpha \in A}\}$ , and let  $U$  be the sequence of all prechoice variables. For some  $Y$  and  $X$ , there exist potential outcomes  $Y^X$  in the sense of Definition 6.6 if and only if  $\mathbb{T}^{Y|UX}$  exists and is deterministic.*

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

Only if: By definition,  $Y^X = g \circ U$ . From Lemma 6.5,  $\mathbb{T}^{Y|XY^X}$  exists and is deterministic. Thus  $\mathbb{T}^{Y|XW}$  also exists and is also deterministic.  $\square$

**Corollary 6.9.** *Potential outcomes  $Y^X$  in the sense of Definition 6.6 exist only if*

$$Y \perp\!\!\!\perp_{\mathbb{T}} D|WX \quad (276)$$

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

Note the similarity between Equation 276 and the condition for proxy control in the previous section. Indeed, the two are identical if we identify  $U$  with  $(W_j, V_A, X_j)$ .

## 7 Appendix:see-do model representation

### Update notation

**Theorem 7.1** (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  $X, Y$  and  $D$  such that  $X_{xdy} = x$ ,  $Y_{xdy} = y$  and  $D_{xdy} = d$ . Then making the following additional assumptions:*

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

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

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

$$\mathbb{P}[XDY]^{ijk} = \mathbb{P}[X]^i \mathbb{P}[D|X]_i^j \mathbb{P}[Y|XD]_{ij}^k \quad (277)$$

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

$$\mathbb{T}_{hj}^{kl} = \mathbb{P}_{h,\mathbb{S}}[X]^k \mathbb{P}_{h,\mathbb{S}}[Y|XD]_{kj}^l \quad (278)$$

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

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

We then have

$$\mathbb{T}[\mathbf{X}'|\mathbf{H}'\mathbf{D}']_{hj}^k = \mathbb{T}\mathbf{X}'_{hj}^k \quad (279)$$

$$= \sum_l \mathbb{T}_{hj}^{kl} \quad (280)$$

$$= \mathbb{P}_{h,\mathbb{S}}[\mathbf{X}]^k \quad (281)$$

$$= \mathbb{T}[\mathbf{X}'|\mathbf{H}'\mathbf{D}']_{hj'}^k \quad (282)$$

Thus  $\mathbf{X}' \perp\!\!\!\perp_{\mathbb{T}} \mathbf{D}'|\mathbf{H}'$  and so  $\mathbb{T}[\mathbf{X}'|\mathbf{H}']$  exists (section 2.8.2) and  $(\mathbb{T}, \mathbf{H}', \mathbf{D}', \mathbf{X}', \mathbf{Y}')$  is a see-do model.

Applying Equation 277 to  $\mathbb{P}_{h,\mathbb{S}}$ :

$$\mathbb{P}_{h,\mathbb{S}}[\mathbf{XDY}]^{ijk} = \mathbb{P}_{h,\mathbb{S}}[\mathbf{X}]^i \mathbb{P}_{h,\mathbb{S}}[\mathbf{D}|\mathbf{X}]_i^j \mathbb{P}_{h,\mathbb{S}}[\mathbf{Y}|\mathbf{XD}]_{ij}^k \quad (283)$$

$$= \mathbb{P}_{h,\mathbb{S}}[\mathbf{X}]^i \mathbb{P}_{h,\mathbb{S}}[\mathbf{Y}|\mathbf{XD}]_{ij}^k \quad (284)$$

$$= \mathbb{P}_{h,\mathbb{S}}[\mathbf{D}|\mathbf{X}]_i^j \mathbb{T}[\mathbf{X}'\mathbf{Y}'|\mathbf{H}'\mathbf{D}']_{hj}^{ik} \quad (285)$$

$$= \mathbb{S}_i^j \mathbb{T}[\mathbf{X}'\mathbf{Y}'|\mathbf{H}'\mathbf{D}']_{hj}^{ik} \quad (286)$$

$$= \mathbb{S}_i^j \mathbb{T}[\mathbf{X}'|\mathbf{H}'\mathbf{D}']_{hj}^i \mathbb{T}[\mathbf{Y}'|\mathbf{X}'\mathbf{H}'\mathbf{D}']_{ihj}^k \quad (287)$$

$$= \mathbb{T}[\mathbf{X}'|\mathbf{H}']_h^i \mathbb{S}_i^j \mathbb{T}[\mathbf{Y}'|\mathbf{X}'\mathbf{H}'\mathbf{D}']_{ihj}^k \quad (288)$$

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

$$\mathbb{P}_{h,\mathbb{S}}[\mathbf{X}]^i = \mathbb{P}_{h,\mathbb{Q}}[\mathbf{X}]^i \quad (289)$$

$$\mathbb{P}_{h,\mathbb{S}}[\mathbf{Y}|\mathbf{XD}]_{ij}^k = \mathbb{P}_{h,\mathbb{Q}}[\mathbf{Y}|\mathbf{XD}]_{ij}^k \text{ for some version of } \mathbb{P}_{h,\mathbb{Q}}[\mathbf{Y}|\mathbf{XD}] \quad (290)$$

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

$$\mathbb{T}_{hj}^{kl} = \mathbb{P}_{h,\mathbb{Q}}[\mathbf{X}]^k \mathbb{P}_{h,\mathbb{Q}}[\mathbf{Y}|\mathbf{XD}]_{kj}^l \quad (291)$$

Then by substitution of  $\mathbb{Q}$  for  $\mathbb{S}$  in Equation 283 and working through the same steps

$$\mathbb{P}_{h,\mathbb{S}}[\mathbf{XDY}]^{ijk} = \mathbb{T}[\mathbf{X}'|\mathbf{H}']_h^i \mathbb{Q}_i^j \mathbb{T}[\mathbf{Y}'|\mathbf{X}'\mathbf{H}'\mathbf{D}']_{ihj}^k \quad (292)$$

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

## 8 Appendix: Counterfactual representation

**Definition 8.1** (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  $\mathbb{K}[\mathbf{Y}_i|\mathbf{W}_i\mathbf{Y}(W)]_{kj_1j_2\dots j_{|W|}} = \delta[j_k]$ .

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

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.

## 8.1 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 than

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 8.2** (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 8.3** (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 8.4** (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 8.5** (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 8.6** (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  $y := (y_i)_{i \in |D|} \in Y^{|D|}$ ,  $d := (d_i)_{i \in [n]} \in D^{[n]}$ ,  $x := (x_i)_{i \in [n]} \in Y^{[n]}$ .

$$\mathbb{K}[X_n|D_nZ]_{dz}^x = \sum_{y \in Y^{|D|}} \mathbb{K}[X_{[n]}|D_nZY(D)]_{dzy}^x \mathbb{K}[Y(D)|D_{[n]}Z]_{dz}^y \quad (293)$$

$$= \sum_{y \in Y^{|D|}} \prod_{i \in [n]} \delta[y_{d_i}]^{x_i} \mathbb{K}[Y(D)|D_nZ]_{dz}^y \quad (294)$$

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

$$\mathbb{K}[X_{[n]}|D_{[n]}Z]_{dz}^x = \sum_{y \in Y^{|\mathcal{D}|}} \prod_{i \in Q_d} \delta[y_{d_i}]^{x_i} \mathbb{K}[Y(D)|D_{[n]}Z]_{dz}^y \quad (295)$$

$$= \sum_{y' \in Y^{|\mathcal{R}_d^C|}} \mathbb{K}[Y(R_d)Y(R_d^C)|D_{Q_d}D_{Q_d^C}Z]_{d_{Q_d}d_{Q_d^C}z}^{x_{Q_d}y'} \quad (296)$$

$$= \sum_{y' \in Y^{|\mathcal{R}_d^C|}} \mathbb{K}[Y_{R_d}Y_{R_d^C}|D_{Q_d}D_{Q_d^C}Z]_{dz}^{x_{Q_d}y'} \quad (297)$$

$$= \sum_{y' \in Y^{|\mathcal{R}_d^C|}} \mathbb{K}[Y_{[n]}|D_{[n]}Z]_{dz}^{x_{Q_d}y'} \quad (\text{using exchangeability}) \quad (298)$$

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.  $\square$

## 9 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

$$I_{K..} := I_K \setminus I_L \setminus I_J \quad (299)$$

$$I_{KL.} := I_K \cap I_L \setminus I_J \quad (300)$$

$$I_{K..J} := I_K \cap I_J \setminus I_L \quad (301)$$

$$I_{KLJ} := I_K \cap I_L \cap I_J \quad (302)$$

$$I_{L.} := I_L \setminus I_K \setminus I_J \quad (303)$$

$$I_{LJ} := I_L \cap I_J \setminus I_K \quad (304)$$

$$I_{..J} := I_J \setminus I_K \setminus I_L \quad (305)$$

$$O_{K..} := O_K \setminus I_N \setminus I_J \quad (306)$$

$$O_{KL.} := O_K \cap I_L \setminus I_J \quad (307)$$

$$O_{K..J} := O_K \cap I_J \setminus I_L \quad (308)$$

$$O_{KLJ} := O_K \cap I_L \cap I_J \quad (309)$$

$$O_{L.} := O_L \setminus I_J \quad (310)$$

$$O_{LJ} := O_L \cap I_J \quad (311)$$

Also define

$$(\mathbb{P}, l_P, O_P) := \mathbb{K} \Rightarrow \mathbb{L} \quad (312)$$

$$(\mathbb{Q}, l_Q, O_Q) := \mathbb{L} \Rightarrow \mathbb{J} \quad (313)$$

Then

$$(\mathbb{K} \Rightarrow \mathbb{L}) \Rightarrow \mathbb{J} = \mathbb{P} \Rightarrow \mathbb{J} \quad (314)$$

$$= \begin{array}{c} l_P. \\ l_{PJ} \end{array} \begin{array}{c} \boxed{\mathbb{P}} \\ \bullet \end{array} \begin{array}{c} O_P. \\ O_{PJ} \end{array} \begin{array}{c} l_J. \\ \bullet \end{array} \begin{array}{c} \boxed{\mathbb{J}} \\ \bullet \end{array} O_J \quad (315)$$

$$= \begin{array}{c} l_{K..} \\ l_{KL.} \\ l_{.L.} \\ l_{K.J} \\ l_{KLJ} \\ l_{.LJ} \\ l_{..J} \end{array} \begin{array}{c} \boxed{\mathbb{K}} \\ \bullet \end{array} \begin{array}{c} \boxed{\mathbb{L}} \\ \bullet \end{array} \begin{array}{c} \boxed{\mathbb{J}} \\ \bullet \end{array} \begin{array}{c} O_{K..} \\ O_{KL.} \\ O_{K.J} \\ O_{KLJ} \\ O_{L.} \\ O_{LJ} \\ O_J \end{array} \quad (316)$$

$$\stackrel{perm}{=} \begin{array}{c} l_{K..} \\ l_{KL.} \\ l_{K.J} \\ l_{KLJ} \\ l_{.L.} \\ l_{.LJ} \\ l_{..J} \end{array} \begin{array}{c} \boxed{\mathbb{K}} \\ \bullet \end{array} \begin{array}{c} \boxed{\mathbb{L}} \\ \bullet \end{array} \begin{array}{c} \boxed{\mathbb{J}} \\ \bullet \end{array} \begin{array}{c} O_{K..} \\ O_{KL.} \\ O_{K.J} \\ O_{KLJ} \\ O_{L.} \\ O_{LJ} \\ O_J \end{array} \quad (317)$$

$$= \begin{array}{c} l_{K.} \\ l_{KQ} \end{array} \begin{array}{c} \boxed{\mathbb{K}} \\ \bullet \end{array} \begin{array}{c} O_{K.} \\ O_{KQ} \end{array} \begin{array}{c} l_Q. \\ \bullet \end{array} \begin{array}{c} \boxed{\mathbb{Q}} \\ \bullet \end{array} O_Q \quad (318)$$

$$= \mathbb{K} \Rightarrow (\mathbb{L} \Rightarrow \mathbb{J}) \quad (319)$$

## 10 Appendix: String Diagram Examples

Recall the definition of *connection*:

**Definition 10.1** (Connection).

$$\mathbb{K} \Rightarrow \mathbb{L} := \begin{array}{c} \text{F}^{\cdot} \text{---} \boxed{\mathbb{K}} \text{---} \text{O}^{\cdot} \text{F}^{\cdot} \\ \text{F}_S \text{---} \bullet \text{---} \text{O}_{FS} \\ \text{I}_S \text{---} \text{---} \boxed{\mathbb{L}} \text{---} \text{O}_S \end{array} \quad (320)$$

$$:= \mathbb{J} \tag{321}$$

$$\mathbb{J}_{yqr}^{zxw} = \mathbb{K}_{yq}^{zx} \mathbb{L}_{xqr}^w \quad (322)$$

Equation 320 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:

$$(\mathbb{J}, (\mathbf{l}_{F\cdot}, \mathbf{l}_{FS}, \mathbf{l}_S), (\mathbf{o}_{F\cdot}, \mathbf{o}_{FS}, \mathbf{o}_S)) = \left[ \begin{array}{c} \text{--- } \mathbf{l}_{F\cdot} \\ \text{--- } \mathbf{l}_{FS} \bullet \text{---} \\ \text{--- } \mathbf{l}_S \end{array} \right] \left[ \begin{array}{c} \boxed{\mathbb{K}} \\ \text{---} \\ \text{---} \end{array} \right] \left[ \begin{array}{c} \text{---} \\ \text{---} \bullet \text{---} \\ \text{---} \end{array} \right] \left[ \begin{array}{ccc} & \mathbf{o}_S & \\ \text{---} & \circ_{FS} & \\ \boxed{\mathbb{L}} & \mathbf{o}_F & \end{array} \right] \quad (323)$$

(324)

## 11 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

$$\delta(x, A) = \begin{cases} 1 & x \in A \\ 0 & \text{otherwise} \end{cases} \quad (325)$$

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

**Definition 11.1.** 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

$$\text{del}_X := \text{---} * \text{copy}_X \quad := \text{---} \bullet \text{---} \quad (326)$$

and satisfying the commutative comonoid equations

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \text{---} \bullet \searrow \text{---} \\ \searrow \text{---} \end{array} \\ \text{---} \end{array} = \begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \text{---} \bullet \searrow \text{---} \\ \searrow \text{---} \end{array} \\ \text{---} \end{array} \quad (327)$$

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \text{---} \\ \searrow \text{---} \end{array} \\ \text{---} \end{array} = \text{---} = \begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \text{---} \\ \searrow \text{---} \end{array} \\ \text{---} \end{array} \quad (328)$$

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \text{---} \\ \searrow \text{---} \end{array} \\ \text{---} \end{array} = \begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \text{---} \\ \searrow \text{---} \end{array} \\ \text{---} \end{array} \quad (329)$$

as well as compatibility with the monoidal structure

$$X \otimes Y \text{---} * = X \text{---} * \quad (330)$$

$$X \otimes Y \text{---} \bullet \begin{array}{l} \nearrow X \otimes Y \\ \searrow X \otimes Y \end{array} = \begin{array}{c} X \text{---} \bullet \begin{array}{l} \nearrow X \\ \searrow Y \end{array} \\ Y \text{---} \bullet \begin{array}{l} \nearrow X \\ \searrow Y \end{array} \end{array} \quad (331)$$

and the naturality of  $del$ , which means that

$$\begin{array}{c} \text{---} \boxed{f} \text{---} * \\ \text{---} \end{array} = \text{---} * \quad (332)$$

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

$$\mathbb{K}\mathbb{L} : X \rightarrow \Delta(Z) \quad (333)$$

$$\text{defined by } (\mathbb{K}\mathbb{L})(x, A) = \int_Y \mathbb{L}(y, A) \mathbb{K}(x, dy) \quad (334)$$

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

$$\mathbb{K} \otimes \mathbb{L} : (X, W) \rightarrow \Delta(Y, Z) \quad (335)$$

$$\text{defined by } \mathbb{K} \otimes \mathbb{L}(x, w, A \times B) = \mathbb{K}(x, A) \mathbb{L}(w, B) \quad (336)$$

The identity map is

$$\text{Id}_X : X \rightarrow \Delta(X) \quad (337)$$

$$\text{defined by } (\text{Id}_X)(x, A) = \delta(x, A) \quad (338)$$

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

$$\text{swap}_{X,Y} : (X, Y) \rightarrow \Delta(Y, X) \quad (339)$$

$$\text{defined by } (\text{swap}_{X,Y})(x, y, A \times B) = \delta(x, B)\delta(y, A) \quad (340)$$

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  $(X, (Y, Z)) \cong ((X, Y), Z)$ .

The copy map is given by

$$\text{copy}_X : X \rightarrow \Delta(X, X) \quad (341)$$

$$\text{defined by } (\text{copy}_X)(x, A \times B) = \delta_x(A)\delta_x(B) \quad (342)$$

and the erase map by

$$\text{del}_X : X \rightarrow \Delta(*) \quad (343)$$

$$\text{defined by } (\text{del}_X)(x, A) = \delta(*, A) \quad (344)$$

$$(345)$$

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 11.2** (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 327 to 332 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

- The Basic Symmetries. In Olav Kallenberg, editor, *Probabilistic Symmetries and Invariance Principles*, Probability and Its Applications, pages 24–68. Springer, New York, NY, 2005. ISBN 978-0-387-28861-1. doi: 10.1007/0-387-28861-9\_2. URL [https://doi.org/10.1007/0-387-28861-9\\_2](https://doi.org/10.1007/0-387-28861-9_2).
- Ethan D. Bolker. Functions Resembling Quotients of Measures. *Transactions of the American Mathematical Society*, 124(2):292–312, 1966. ISSN 0002-9947. doi: 10.2307/1994401. URL <https://www.jstor.org/stable/1994401>. Publisher: American Mathematical Society.
- G. Chiribella, Giacomo D’Ariano, and P. Perinotti. Quantum Circuit Architecture. *Physical review letters*, 101:060401, September 2008. doi: 10.1103/PhysRevLett.101.060401.
- Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, August 2019. ISSN 0960-1295, 1469-8072. doi: 10.1017/S0960129518000488.
- Panayiota Constantinou and A. Philip Dawid. EXTENDED CONDITIONAL INDEPENDENCE AND APPLICATIONS IN CAUSAL INFERENCE. *The Annals of Statistics*, 45(6):2618–2653, 2017. ISSN 0090-5364. URL <http://www.jstor.org/stable/26362953>. Publisher: Institute of Mathematical Statistics.
- A. Philip Dawid. Decision-theoretic foundations for statistical causality. *arXiv:2004.12493 [math, stat]*, April 2020. URL <http://arxiv.org/abs/2004.12493>. arXiv: 2004.12493.
- M. P. Ershov. Extension of Measures and Stochastic Equations. *Theory of Probability & Its Applications*, 19(3):431–444, June 1975. ISSN 0040-585X. doi: 10.1137/1119053. URL <https://epubs.siam.org/doi/abs/10.1137/1119053>. Publisher: Society for Industrial and Applied Mathematics.
- R.P. Feynman. *The Feynman lectures on physics*. Le cours de physique de Feynman. Intereditions, France, 1979. ISBN 978-2-7296-0030-3. INIS Reference Number: 13648743.
- Brendan Fong. Causal Theories: A Categorical Perspective on Bayesian Networks. *arXiv:1301.6201 [math]*, January 2013. URL <http://arxiv.org/abs/1301.6201>. arXiv: 1301.6201.
- Tobias Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, August 2020. ISSN 0001-8708. doi: 10.1016/j.aim.2020.107239. URL <https://www.sciencedirect.com/science/article/pii/S0001870820302656>.
- SANDER GREENLAND and JAMES M ROBINS. Identifiability, Exchangeability, and Epidemiological Confounding. *International Journal of Epidemiology*,

- 15(3):413–419, September 1986. ISSN 0300-5771. doi: 10.1093/ije/15.3.413. URL <https://doi.org/10.1093/ije/15.3.413>.
- 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>.
- Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer Science & Business Media, October 2004. ISBN 978-3-540-22139-5. Google-Books-ID: ziLLYu7oIkQC.
- 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 Mikołaj 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.



- 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](https://doi.org/10.1007/978-3-7091-6045-9_31).
- Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009.
- Judea Pearl. Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes. *Journal of Causal Inference*, 6(2), 2018. doi: 10.1515/jci-2018-2001. URL <https://www.degruyter.com/view/j/jci.2018.6.issue-2/jci-2018-2001/jci-2018-2001.xml>.
- Jonas Peters and Peter Bühlmann. Structural Intervention Distance for Evaluating Causal Graphs. *Neural Computation*, 27(3):771–799, January 2015. ISSN 0899-7667. doi: 10.1162/NECO\_a\_00708. URL [https://doi.org/10.1162/NECO\\_a\\_00708](https://doi.org/10.1162/NECO_a_00708).
- 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](https://doi.org/10.1007/978-3-319-20451-2_3).
- Donald B. Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, March 2005. ISSN 0162-1459. doi: 10.1198/016214504000001880. URL <https://doi.org/10.1198/016214504000001880>.
- Leonard J. Savage. *The Foundations of Statistics*. Wiley Publications in Statistics, 1954.
- Peter Selinger. A survey of graphical languages for monoidal categories. *arXiv:0908.3347 [math]*, 813:289–355, 2010. doi: 10.1007/978-3-642-12821-9\_4. URL <http://arxiv.org/abs/0908.3347>. arXiv: 0908.3347.
- Eyal Shahar. The association of body mass index with health outcomes: causal, inconsistent, or confounded? *American Journal of Epidemiology*, 170(8):957–958, October 2009. ISSN 1476-6256. doi: 10.1093/aje/kwp292.
- Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, July 2002.
- J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Theory of games and economic behavior. Princeton University Press, Princeton, NJ, US, 1944.

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

## Appendix: