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

David Johnston

December 30, 2021

Contents

1 Introduction

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

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

For example, Hernán and Taubman (2008), who observed that many epidemiological papers have been published estimating the "causal effect" of body mass index and argued that, because *actions* affecting body mass index[1] are vaguely defined, potential outcome variables and causal effects themselves become ill-defined. We note that "actions targeting body mass index" are not elements of a potential outcomes model but "things to which potential outcomes should correspond". The authors claim is that vagueness in the "territory" leads to ambiguity about elements of the "map" – and, as we have suggested, anything we can try to say about the territory is unavoidably vague. This seems like a serious problem.

In a response, Pearl (2018) argues that *interventions* (by which we mean the operation defined by a causal graphical model) are well defined, but may not always be a good model of an action. Pearl further suggests that interventions in graphical models correspond to "virtual interventions" or "ideal, atomic interventions", and that perhaps carefully chosen interventions can be good models of actions. Shahar (2009), also in response, argued that interventions targeting body mass index applied to correctly specified graphical causal models will necessarily yield no effect on anything else which, together with Pearl's suggestion,

[1] the authors use the term "intervention", but they do not use it mean a formal operation on a graphical causal model, and we reserve the term for such operations to reduce ambiguity.

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 probabilstic reasoning in a number of places, and this section is intended to help interpret mathematical statements in this form.

The second subsection discusses the interpretation of probabilistic variables. Our formalisation of probabilistic variables is standard – we define them as measurable functions on a fundamental probability set Ω. We discusses 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 $\mathbb{P}^{\mathsf{X}}$ for some variable $\mathsf{X} : \Omega \to X$. Such a probability distribution is consistent with many base measures on the fundamental probability set Ω, and so we can consider the choice of base measure to be a probability gap. Not every probability distribution over X can define a probability gap model in this way. In particular, we need $\mathbb{P}^{\mathsf{X}}$ 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 $\mathbb{P}^{\mathsf{X}}$ 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 $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ or pairs of conditional probabilities $\{\mathbb{P}^{\mathsf{X}|\mathsf{W}}, \mathbb{P}^{\mathsf{Z}|\mathsf{WXY}}\}$, which we call *order 1* and *order 2* models respectively. Decision functions in data-driven decision problems correspond to probability gaps in order 2 models, as we discuss in Section 3, which makes this type of model particularly interesting for our purposes. We also require these to be valid, and we define conditions for validity and prove that they are sufficient to ensure that models represented by conditional probabilities can in fact be mapped to base measures on the fundamental probability set.

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

We can consider causal Bayesian networks to represent order 2 probability gap models. That is, a causal Bayesian network represents a function $\mathbb{P}$ that take 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 $\{\mathbb{P}^{\mathsf{X}|\mathsf{W}}, \mathbb{P}^{\mathsf{Z}|\mathsf{WXY}}\}$ 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 Semantics of observed and unobserved variables

We are interested in constructing *probabilistic models* which explain some part of the world. In a model, variables play the role of "pointing to the parts of the world the model is explaining". Both observed an unobserved variables play important roles in causal modelling and we think it is worth clarifying what variables of either type refer to. Our approach is a standard one: a probabilistic model is associated with an experiment or measurement procedure that yields values in a well-defined set. Observable variables are obtained by applying well-defined functions to the result of this total measurement. We use a richer 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 perculiar. Acknowledging that physicists happen to know a great deal about forces on an object, it remains true that in order to actually say what the net force on a real object is, even a highly knowledgeable physicist will still have to go and do some measurements, and the result of such measurements will be a vector representing the net forces on that object.

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

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 2.1 (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 the procedure $\mathcal{B}$ will yield the value $x \in X$. $\mathcal{B} \bowtie A$ for $A \subset X$ is the proposition $\vee_{x\in A}\mathcal{B} \bowtie x$. Two procedures $\mathcal{B}$ and $\mathcal{C}$ are the same if $\mathcal{B} \bowtie x \iff \mathcal{C} \bowtie x$ for all $x \in B$ (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 2.2 (Composition of functions with procedures)**.** Given a procedure $\mathcal{B}$ that takes values in some set B, and a function $f : B \to C$, define the "composition" $f \circ \mathcal{B}$ to be 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, $\mathcal{M}\mathcal{A}$ is the composition of $h : (x, y) \mapsto xy$ with the procedure $(\mathcal{M}, \mathcal{A})$ that yields the mass and acceleration of the same object. Measurement procedure composition is associative:

$$
\begin{aligned}
(g \circ f) \circ \mathcal{B} \text{ yields } x &\iff B \text{ yields } (g \circ f)^{-1}(x) && (1)\\
&\iff B \text{ yields } f^{-1}(g^{-1}(x)) && (2)\\
&\iff f \circ B \text{ yields } g^{-1}(x) && (3)\\
&\iff g \circ (f \circ B) \text{ yields } x && (4)
\end{aligned}
$$

One might whether there is also some kind of "append" operation that takes a standalone $\mathcal{M}$ and a standalone $\mathcal{A}$ and returns a procedure $(\mathcal{M}, \mathcal{A})$. Unlike function composition, this would be an operation that acts on two procedures rather than a procedure and a function. Unlike composition, we can't easily reason about such an operation mathematically, because of the fact that measurment 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 Ψ 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 $\mathcal{F} = \mathcal{M}\mathcal{A}$, for example, we could assume $\mathcal{F} = f \circ \mathcal{S}$ for some f and $(\mathcal{M}, \mathcal{A}) = g \circ \mathcal{S}$ for some g. In this case, we can get $\mathcal{M}\mathcal{A} = h \circ (\mathcal{M}, \mathcal{A}) =$

$(h \circ g) \circ \mathcal{S}$. Note that each procedure is associated with a unique function with domain Ψ.

Given that measurement processes are in practice finite precision and with finite range, Ψ will generally be a finite set. We can therefore equip Ψ with the collection of measurable sets given by the power set $\mathcal{E} := \mathcal{P}(\Psi)$, and $(\Psi, \mathcal{E})$ is a standard measurable space. $\mathcal{E}$ stands for a complete collection of logical propositions we can generate that depend on the results yielded by the measurement procedure $\mathcal{S}$.

$(\Psi, \mathcal{E})$ defines is a "sample space" limited to observable variables. That is, Ψ is associated with a measurement procedure. Unobserved variables need not be associated with measurement procedures, and to accommodate these we use instead of Ψ a richer fundamental probability set Ω which represents both observed and unobserved variables.

Definition 2.3 (Sample space)**.** The sample space $(\Omega, \mathcal{F})$ is a set Ω along with with a σ-algebra $\mathcal{F}$ of subsets of Ω.

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

Definition 2.4 (Consistency with observation)**.** An element $\omega \in \Omega$ is is *consistent with observation* if the result yielded by $\mathcal{S} \bowtie \mathsf{S}(\omega)$

Thus the procedure $\mathcal{S}$ restricts the observationally consistent elements of Ω. If $\mathcal{S}$ yield the result s, then the consistent values of Ω will be $\mathsf{S}^{-1}(s)$. While two different sets of measurement outcomes Ψ and Ψ' entail a different mesurement procedures $\mathcal{S}$ and $\mathcal{S}'$, but different fundamental probability sets Ω and Ω' 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 Ω that do not factorise through $\mathcal{S}$, and these variables do not model procedures.

2.3 Events

To recap, we have a procedure $\mathcal{S}$ yielding values in Ψ that measures everything we are interested in, a fundamental probability set Ω and a function S that models $\mathcal{S}$ in the sense of Definition 2.4. We assume also that Ψ has a σ-algebra $\mathcal{E}$ (this may be the power set of Ψ, as measurement procedures are typically limited to finite precision). Ω is equipped with a σ-algebra $\mathcal{F}$ such that $\sigma(\mathsf{S}) \subset \mathcal{F}$. If a procedure $\mathcal{X} = f \circ \mathsf{S}$ then we define $\mathsf{X} : \Omega \to X$ by $\mathsf{X} := f \circ \mathsf{S}$.

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

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

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

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

2.4 Standard probability theory

Definition 2.5 (Probability measure)**.** Given a measure space $(X, \mathcal{X})$, a probability measure is a σ-additive function $\mu : \mathcal{X} \to [0, 1]$ such that $\mu(\emptyset) = 0$ and $\mu(X) = 1$. We write $\Delta(X)$ for the set of all probability measures on $(X, \mathcal{X})$.

Definition 2.6 (Markov kernel)**.** Given measure spaces $(X, \mathcal{X})$, $(Y, \mathcal{Y})$ $\mathsf{Y} : \Omega \to Y$, a Markov kernel $\mathbb{Q} : X \rightarrowtriangle Y$ is a map $Y \times \mathcal{X} \to [0, 1]$ such that

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

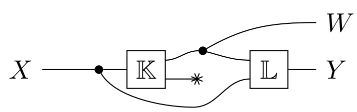
Definition 2.7 (Probability measures as Markov kernel)**.** Given $(X, \mathcal{X})$ and $\mu \in \Delta(X)$, the Markov kernel $\mathbb{K} : \{*\} \rightarrowtriangle X$ given by $\mathbb{K}(A|*) = \mu(A)$ for all $A \in \mathcal{X}$ is the Markov kernel associated with the probability measure μ. We will use probability measures and their associated Markov kernels interchangeably, as it is transparent how to get from one to another.

Definition 2.8 (Delta measure)**.** Given a measure space $(X, \mathcal{X})$ and $x \in X$, $\delta_x \in \Delta(X)$ is the measure defined by $\delta_x(A) = [\![x \in A]\!]$.

Definition 2.9 (Probability space)**.** A probability space is a triple $(\mu, \Omega, \mathcal{F})$, where μ is a base measure on $\mathcal{F}$.

Definition 2.10 (Marginal distribution with respect to a probability space)**.** Given a probability space $(\mu, \Omega, \mathcal{F})$ and a random variable $\mathsf{X} : \Omega \to (X, \mathcal{X})$, we can define the *marginal distribution* of X with respect to μ, $\mu^{\mathsf{X}} : \mathcal{X} \to [0, 1]$ by $\mu^{\mathsf{X}}(A) := \mu(\mathsf{X} \bowtie A)$ for any $A \in \mathcal{X}$.

Definition 2.11 (Disintegration)**.** Given a Markov kernel $\mathbb{K} : W \rightarrow X \times Y$, with W, X and Y standard measurable, any kernel $\mathbb{L} : W \times X \rightarrow Y$ satisfying



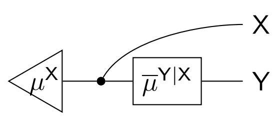
$$\mathbb{K} = \tag{5}$$

is a $W \times X \rightarrow Y$ *disintegration* of $\mathbb{K}$.

Definition 2.12 (Conditional probability with respect to a probability space)**.** Given a probability space (μ, Ω) and random variables $\mathsf{X} : \Omega \to (X, \mathcal{X})$, $\mathsf{Y} : \Omega \to (Y, \mathcal{Y})$, the probability of Y given X is any $X \rightarrow Y$ disintegration of μ^{XY}. That is,

$$\int_A \mu^{\mathsf{Y}|\mathsf{X}}(B|x)\mathrm{d}\mu^{\mathsf{X}}(x) = \mu^{\mathsf{XY}}(x, y) \qquad \forall A \in \mathcal{X}, B \in \mathcal{Y} \tag{6}$$

$$\iff \tag{7}$$



$$= \mu^{\mathsf{XY}} \tag{8}$$

Lemma 2.13 (Marginal distribution as a kernel product)**.** *Given a probability space* $(\mu, \Omega, \mathcal{F})$ *and a random variable* $\mathsf{X} : \Omega \to (X, \mathcal{X})$*, define* $\mathbb{F}_{\mathsf{X}} : \Omega \rightarrow X$ *by* $\mathbb{F}_{\mathsf{X}}(A|\omega) = \delta_{\mathsf{X}(\omega)}(A)$*, then*

$$\mu^{\mathsf{X}} = \mu\mathbb{F}_{\mathsf{X}} \tag{9}$$

Proof. Consier any $A \in \mathcal{X}$.

$$\mu\mathbb{F}_{\mathsf{X}}(A) = \int_\Omega \delta_{\mathsf{X}(\omega)}(A)\mathrm{d}\mu(\omega) \tag{10}$$

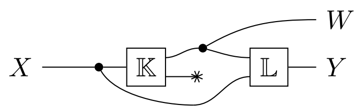
$$= \int_{\mathsf{X}^{-1}(\omega)} \mathrm{d}\mu(\omega) \tag{11}$$

$$= \mu^{\mathsf{X}}(A) \tag{12}$$

□

Disintegration of arbitrary Markov kernels is possible in standard measurable spaces. We will assume that all spaces are standard measurable, such that whenever we have a Markov kernel we are able to disintegrate it.

Lemma 2.14 (Disintegration existence in standard measurable Markov kernels)**.** *For any Markov kernel $\mathbb{K} : X \rightarrowtriangle W \times Y$ and X, W, Y are standard measurable, there exists $\mathbb{L} : W \times X \rightarrowtriangle Y$ such that*



$$\mathbb{K} = \tag{13}$$

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

Proof. Cho and Jacobs (2019) Theorem 3.11 □

2.5 Probabilistic models for causal inference

The sample space $(\Omega, \mathcal{F})$ along with our collection of variables is a "model skeleton" – it tells us what kind of data we might see. The process $\mathcal{S}$ which tells us which part of the world we're interested in is related to the model Ω and the observable variables by the criterion of *consistency with observation*. The kind of problem we are mainly interested in here is one where we make use of data to help make decisions under uncertainty. Probabilistic models have a long history of being used for this purpose, and our interest here is in constructing probabilistic models that can be attached to our variable "skeleton".

Given a model skeleton, a common approach to attaching a probabilistic model involves defining a base measure μ on $(\Omega, \mathcal{F})$ which yields a probability space $(\Omega, \mathcal{F}, \mu)$. For causal inference, we need a to generalise this approach, because we need to handle *choices*. If I have different options I can choose, and I want to use a model to compare the options according to some criteria, then I need a model that can accept a choice and output the expected result of that choice. According to this model, anything that we consider a "consequence of a choice" doesn't have a definite probability, because it depends on the choice we make.

In general, we might have arbitrary sets of choices that map to probability models in an arbitrary way. However, we are here interested in a simpler case: we suppose that there are a number of points at which we can act, and prior to acting we can observe some variables, and we are able to choose probabilistic maps from observations to acts. We also assume that, given the same observation and the same act, the same consequence is expected. That is, the consequences do not depend directly way on the choice of map from observations to acts.

These assumptions together imply that our model should contain a number of fixed conditional probabilities – the probabilities of consequences given observations and acts – and a number of "choosable" conditional probabilities – the probabilities of acts given observations. The fixed conditional probabilities form a probability model with *gaps*, and those gaps correspond to choices we can make. When we combine the fixed conditional probabilities and a choice of a conditional probability for each gap, we get a regular probability model. The terminology of "probability gaps" comes from Hájek (2003).

To restate our general approach: we model decision problems with a collection of fixed conditional probabilities and a collection of choosable conditional probabilities, and combine the fixed conditionals with particular choices to get a probability measure. Two issues present themselves here: firstly, what *is* a collection of conditional probabilities without a fixed underlying probability measure? Secondly, we need to ensure that our chosen collection of conditional probabilities actually does induce a probability model. We address these questions with *probability sets*. A probability set is a collection of probability measures on $(\Omega, \mathcal{F})$, and we identify a collection of conditional probabilities with the set of probability measures that induce those conditional probabilities. We then define an operation $\odot$ for combining conditional probabilities, and a criterion of *validity* such that a collection of valid conditional probabilities recursively combined using $\odot$ is guaranteed to corresponds to a non-empty probability set.

2.6 Probability sets

A probability set is a set of probability measures. This section establishes a number of useful properties of conditional probability with respect to probability sets. Unlike conditional probability with repsect to a probability space, conditional probabilities don't always exist for probability sets. Where they do, however, they are almost surely unique and we can marginalise and disintegrate them to obtain other conditional probabilities with respect to the same probability set.

Definition 2.15 (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 $\mathcal{P}(\Delta(\Omega))$, where $\mathcal{P}$ indicates the power set.

Given a probability set $\mathbb{P}_{\{\}}$, we define marginal and conditional probabilities as probability measures and Markov kernels that satisfy Definitions 2.10 and 2.12 respectively for *all* base measures in $\mathbb{P}_{\{\}}$. There are generally multiple Markov kernels that satisfy the properties of a conditional probability with respect to a probability set, and this definition ensures that marginal and conditional probabilities are "almost surely" unique (Definition 2.21) with respect to probability sets.

Definition 2.16 (Marginal probability with respect to a probability set)**.** Given a sample space $(\Omega, \mathcal{F})$, a variable $\mathsf{X} : \Omega \to X$ and a probability set $\mathbb{P}_{\{\}}$, the marginal distribution $\mathbb{P}^{\mathsf{X}}_{\{\}} = \mathbb{P}^{\mathsf{X}}_{a}$ for any $\mathbb{P}_a \in \mathbb{P}_{\{\}}$ if a distribution satisfying this condition exists. Otherwise, it is undefined.

Definition 2.17 (Conditional probability with respect to a probability set)**.** Given a fundamental probability set Ω variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$ and a probability set $\mathbb{P}_{\{\}}$, *a version of* $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ is any Markov kernel $X \rightarrowtriangle Y$ such that $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ is an $X \rightarrowtriangle Y$ disintegration of $\mathbb{P}_a^{\mathsf{XY}}$ for all $\mathbb{P}_a \in \mathbb{P}_{\{\}}$. If no such Markov kernel exists, $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ is undefined.

Given a conditional probability with respect to a probability set, we can find other conditional probabilities by "pushing it forward".

Theorem 2.18 (Recursive pushforward)**.** *Suppose we have a sample space Ω variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_{\{\}}$ such that $\mathbb{P}_{\{\}}^{\mathsf{X}|\mathsf{Y}}$ is a $\mathsf{Y}|\mathsf{X}$ conditional probability of $\mathbb{P}_{\{\}}$ and $\mathsf{Z} = f \circ \mathsf{Y}$ for some $f : Y \to Z$. Then there exists a $\mathsf{Z}|\mathsf{X}$ conditional probability of $\mathbb{P}_{\{\}}$ given by $\mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{X}} = \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \mathbb{F}_f$.*

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

$$\mathbb{P}_a^{\mathsf{X}}(x)\mathbb{P}_a^{\mathsf{Z}|\mathsf{X}}(z|x) = \mathbb{P}_a(\mathsf{X}^{-1}(x) \cap \mathsf{Z}^{-1}(z)) \tag{14}$$

$$= \mathbb{P}_a(\mathsf{X}^{-1}(x) \cap \mathsf{Y}^{-1}(f^{-1}(z))) \tag{15}$$

$$= \mathbb{P}_a^{\mathsf{X},\mathsf{Y}}(\{x\} \times f^{-1}(z)) \tag{16}$$

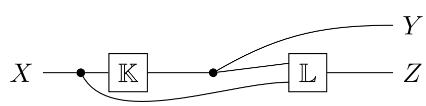
$$= \mathbb{P}_a^{\mathsf{X}}(x)\mathbb{P}_a^{\mathsf{Y}|\mathsf{X}}(f^{-1}(z)|x) \tag{17}$$

□

We define the copy-product $\odot$ as a shorthand for the operation in Equation **??** that combines a marginal with a disintegration to get the original Markov kernel back.

Definition 2.19 (Copy product)**.** Given $\mathbb{K} : X \rightarrowtriangle Y$ and $\mathbb{L} : Y \times X \rightarrowtriangle Z$, define the copy-product $\mathbb{K} \odot \mathbb{L} : X \to Y \times Z$ as

$$\mathbb{K} \odot \mathbb{L} := \mathrm{copy}_X(\mathbb{K} \otimes \mathrm{id}_X)(\mathrm{copy}_Y \otimes \mathrm{id}_X)(\mathrm{id}_Y \otimes \mathbb{L}) \tag{18}$$

$=$  (19)

$$\iff \tag{20}$$

$$(\mathbb{K} \odot \mathbb{L})(A \times B|x) = \int_A \mathbb{L}(B|y,x)\mathbb{K}(dy|x) \qquad A \in \mathcal{Y}, B \in \mathcal{Z} \tag{21}$$

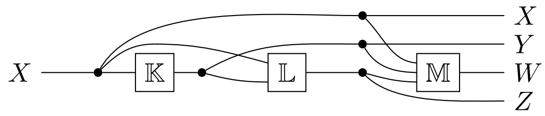
Lemma 2.20 (Copy product is associative)**.** *Given $\mathbb{K} : X \rightarrowtriangle Y$, $\mathbb{L} : Y \times X \rightarrowtriangle Z$ and $\mathbb{M} : Z \times Y \times X \rightarrowtriangle W$*

$$(\mathbb{K} \odot \mathbb{L}) \odot \mathbb{Z} = \mathbb{K} \odot (\mathbb{L} \odot \mathbb{Z}) \tag{22}$$

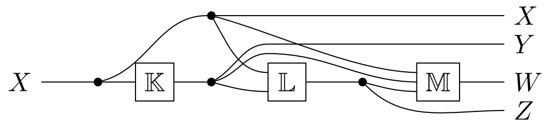
(23)

Proof.

$$(\mathbb{K} \odot \mathbb{L}) \odot \mathbb{M} = \tag{24}$$



$$= \tag{25}$$



$$= \mathbb{K} \odot (\mathbb{L} \odot \mathbb{M}) \tag{26}$$

□

Two Markov kernels are almost surely equal with respect to a probability set $\mathbb{P}_{\{\}}$ if the copy product $\odot$ of all marginal probabilities of X with each Markov kernel is identical.

Definition 2.21 (Almost sure equality)**.** Two Markov kernels $\mathbb{K} : X \to Y$ and $\mathbb{L} : X \to Y$ are almost surely equal $\overset{\mathbb{P}_{\{\}}}{\cong}$ with respect to a probability set $\mathbb{P}_{\{\}}$ and variable $\mathsf{X} : \Omega \to X$ if for all $\mathbb{P}_a \in \mathbb{P}_{\{\}}$,

$$\mathbb{P}_a^{\mathsf{X}} \odot \mathbb{K} = \mathbb{P}_a^{\mathsf{X}} \odot \mathbb{L} \tag{27}$$

Lemma 2.22 (Conditional probabilities are almost surely equal)**.** *If* $\mathbb{K} : X \to Y$ *and* $\mathbb{L} : X \to Y$ *are both versions of* $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$, $\mathbb{K} \overset{\mathbb{P}_{\{\}}}{\cong} \mathbb{L}$

Proof. For all $\mathbb{P}_a \in \mathbb{P}_{\{\}}$

$$\mathbb{P}_a^{\mathsf{X}} \odot \mathbb{K} = \mathbb{P}_a^{\mathsf{XY}} \tag{28}$$

$$= \mathbb{P}_a^{\mathsf{X}} \odot \mathbb{L} \tag{29}$$

□

Lemma 2.23 (Substitution of almost surely equal Markov kernels)**.** *Given* $\mathbb{P}_{\{\}}$, *if* $\mathbb{K} : X \times Y \to Z$ *and* $\mathbb{L} : X \times Y \to Z$ *are almost surely equal* $\mathbb{K} \overset{\mathbb{P}_{\{\}}}{\cong} \mathbb{L}$, *then for any* $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{K} \overset{a.s.}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{L} \tag{30}$$

Proof. For any $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$

$$\mathbb{P}_\alpha^{\mathsf{XY}} \odot \mathbb{K} = (\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}) \odot \mathbb{K} \tag{31}$$

$$= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{K}) \tag{32}$$

$$= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{L}) \tag{33}$$

□

Lemma 2.24 (Copy product of conditionals is a joint conditional)**.** *Given a probability set* $\mathbb{P}_{\{\}}$ *on* $(\Omega, \mathcal{F})$ *along with conditional probabilities* $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ *and* $\mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{XY}}$*,* $\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$ *exists and is equal to*

$$\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}} = \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{XY}} \tag{34}$$

$$\tag{35}$$

Proof. By definition, for any $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$

$$\mathbb{P}_\alpha^{\mathsf{XYZ}} = \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_\alpha^{\mathsf{YZ}|\mathsf{X}} \tag{36}$$

$$= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_\alpha^{\mathsf{Z}|\mathsf{YX}}) \tag{37}$$

$$= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}) \tag{38}$$

□

2.7 Probability sets defined by marginal and conditional probabilities

In the previous section we defined marginal and conditional probabilities for probability sets. Here we will go in the other direction: define probability sets by specifying key marginal or conditional probabilities. There is an issue to be careful of here: not all probability measures $\mathbb{Q}^{\mathsf{X}}$ on X define nonempty sets of probability measures on Ω with respect to the variable X. Consider, for example, $\mathsf{X} = (\mathsf{Y}, \mathsf{Y})$ for some $\mathsf{Y} : \Omega \to \{0, 1\}$ and any measure $\mathbb{Q}^{\mathsf{YY}}$ that assigns nonzero probability to the event $(\mathsf{Y}, \mathsf{Y}) \bowtie (1, 0)$. There is no base measure that pushes forward to such a $\mathbb{P}^{\mathsf{YY}}$, because two copies of the same variable must always be deterministically equal. A *valid distribution* is a distribution associated with a particular variable that defines a nonempty set of base measures on Ω (Theorem 4.15).

Definition 2.25 (Valid distribution)**.** A valid X probability distribution $\mathbb{P}^{\mathsf{X}}$ is any probability mesure on $\Delta(X)$ such that $\mathsf{X}^{-1}(A) = \emptyset \implies \mathbb{P}^{\mathsf{X}}(A) = 0$ for all $A \in \mathcal{X}$.

Valid conditionals not only define a nonempty set of base measures on Ω, but the $\odot$ product of two appropriately typed valid conditionals itself defines a nonempty set of base measures on Ω (Lemma 4.18).

Definition 2.26 (Valid conditional)**.** Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ a *valid* $\mathsf{Y}|\mathsf{X}$ *conditional probability* $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ is a Markov kernel $X \rightarrow Y$ such that it assigns probability 0 to contradictions:

$$\forall B \in \mathcal{Y}, x \in \mathcal{X} : (\mathsf{X}, \mathsf{Y}) \bowtie \{x\} \times B = \emptyset \implies \left(\mathbb{P}^{\mathsf{Y}|\mathsf{X}}(A|x) = 0\right) \vee (\mathsf{X} \bowtie \{x\} = \emptyset) \tag{39}$$

Thus, given a collection of valid conditional probabilities $\{\mathbb{P}_i^{\mathsf{X}_i|\mathsf{X}_{[i-1]}}|i \in [n]\}$ such that each adjacent pair can be combined with the $\odot$ product, the sequential product of each conditional probability is a valid conditional probability and there is a non-empty set of probability measures on the sample space that with that conditional probability. Collections of recursive conditional probabilities often arise in causal modelling – in particular, they are the foundation of the structural equation modelling approach Richardson and Robins (2013); Pearl (2009). Lemma 4.18 establishes that recursive collections of conditional probabilities define non-empty probability sets as long as all the conditional probabilities in the collection are valid.

Definition 2.27 (Probability set defined by a valid conditional)**.** Given a valid conditional probability $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} : X \rightarrowtriangle Y$, let $\mathbb{P}_{\{\}}$ be the probability set containing every $\mu \in \Delta(\Omega)$ such that $\mu^{\mathsf{Y}|\mathsf{X}} \overset{\mu}{\cong} \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$.

We're making complicated models where we can choose many different combinations of conditional probabilities, validity is useful because we can check just the individual conditionals, don't need to check all the combinations. Also, there are examples in the wild of invalid collections – see BMI example

While structural equation models posit a complete collections of conditional probabilities with an operation, often called "intervention", that replaces some of them, we are interested in working with incomplete collections with a domain of choices that complete the collection; we call such objects "probability gap models". Structural equation models are like probability gap models with default choices – if you don't replace anything, you get the basic set of structural equations, but you can also choose different conditionals from some domain.

2.8 Probability combs

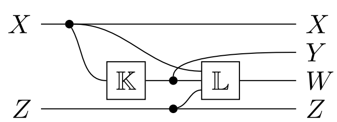
We are interested in a a more precise kind of of "incomplete collection of conditional probabilities", and the specific kind of thing we are interested in is a *probability comb*. Probability combs can be represented either as a collection of conditional probabilities with every second element missing – the "disassembled" representation – or as a single Markov kernel – the "assembled" representation. The key result of this section is the equivalence of the two representations and the fact that assembled probability combs are conditional probabilities with respect to *blind* choices (that is, choices for which the output does not depend on the input). This means that to specify a probability comb, under some conditions it is enough to specify the conditional probability with respect to a blind choice.

In the following, D is an index set which is equal to either $\mathbb{N}$ (if it is infinite) or $[n]$ for some $n \in \mathbb{N}$ (if it is finite). The indexing of the set of variables is arbitrary, it is just a means of addressing each variable. D_{even} are the even elements of D and D_{odd} are the odd elements.

Definition 2.28 (Disassembled probability comb)**.** Given sample space $(\Omega, \mathcal{F})$ and a collection of $|D|$ variables $\mathsf{X}_i : \Omega \to (X_i, \mathcal{X}_i)$ for $i \in D \cup \{0\}$. A $\mathsf{X}_{D_{\text{odd}}} \wr \mathsf{X}_{D_{\text{even}}}$ *disassembled probability comb* is a collection of valid conditional probabilities $\{\mathbb{P}_\square^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in \mathsf{X}_{D_{\text{odd}}}\}$.

To define the assembled form, it is helpful to define a new operation $\underline{\odot}$. Note that consecutive conditional probabilities in a disassembled probability comb have the form $\mathbb{P}_\square^{\mathsf{X}_i|\mathsf{X}_{[i-1]}}, \mathbb{P}_\square^{\mathsf{X}_{i+2}|\mathsf{X}_{[i+1]}}$, and so they are not correctly typed to take the $\odot$ product. However, we can take the $\odot$ product of $\mathbb{P}_\square^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} \otimes \mathrm{id}_{X_{i+1}}$ and $\mathbb{P}_\square^{\mathsf{X}_{i+2}|\mathsf{X}_{[i+1]}}$. The operation $\underline{\odot}$ is just this – it takes the first Markov kernel tensored with the identity, then computes the $\odot$ product of the resulting Markov kernels.

Definition 2.29 (Bypass copy product)**.** Given two Markov kernels $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \times Y \times Z \rightarrow W$, $\mathbb{K}\underline{\odot}\mathbb{L} : X \times Z \rightarrow X \times Y \times Z \times W$ is the Markov kernel equal to

$\mathbb{K}\underline{\odot}\mathbb{L} :=$  (40)

We will also use a series notation for a sequence of $\underline{\odot}$ products.

Definition 2.30 (Bypass copy product series notation)**.** For a collection of Markov kernels $\{\mathbb{K}_i | i \in D_{\text{odd}}\}$ with $\mathbb{K}_i : X_{[i-1]} \rightarrow X_i$ and $n \in \mathbb{N}$ a finite index set

$$\underline{\bigodot}_{i \in [n]_{\text{odd}}} \mathbb{K}_i = \mathbb{K}_1 \underline{\odot} (\underline{\bigodot}_{i \in [n]_{\text{odd}} \setminus \{1\}} \mathbb{K}_i) \tag{41}$$

with base case

$$\underline{\bigodot}_{i \in \{x\}} \mathbb{K}_i = \mathbb{K}_x \tag{42}$$

Definition 2.31 (Assembled probability comb)**.** Given sample space $(\Omega, \mathcal{F})$, $n \in \mathbb{N}$, $\mathsf{X}_i : \Omega \to (X_i, \mathcal{X}_i)$ for $i \in [n]$ and a disassembled probability comb $\{\mathbb{P}_\square^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in \mathsf{X}_{[n]_{\text{odd}}}\}$, the assembled probability comb

$$\mathbb{P}_\square^{\mathsf{X}_{[n]_{\text{odd}}} \wr \mathsf{X}_{[n]_{\text{even}}}} = \underline{\bigodot}_{i \in [n]_{\text{odd}}} \mathbb{P}_\square^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} \tag{43}$$

Whether we can extend comb products and the following theorem to infinite collections of conditional probabilities is an open question.

Theorem 2.32 establishes that there is a bijection, up to almost sure equality, between disassembled and assembled comb representations. We can assemble the collection of conditional probabilities as in Definition 2.31 to get the assembled representation, and we can disintegrate an assembled probability comb to get a collection of conditional probabilities. Note that we only show this to hold for discrete sets and finite collections of conditional probabilities.

Theorem 2.32 (Equivalence of comb representations)**.** *Given sample space $(\Omega, \mathcal{F})$, a finite collection of variables $\mathsf{X}_i : \Omega \to (X_i, \mathcal{X}_i)$ for $i \in [n]$, X_i discrete, and a disassembled probability comb $\{\mathbb{P}_\square^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in \mathsf{X}_{[n]_{odd}}\}$, for any $l \in [n]_{\text{odd}}$ and any $\mathbb{K} : X_{[l-1]} \rightarrowtriangle X_l$*

$$(\underline{\bigodot}_{j\in[l-1]_{\text{odd}}} \mathbb{P}_\square^{\mathsf{X}_j|\mathsf{X}_{[j-1]}})\underline{\odot}\mathbb{K} \overset{\mathbb{P}_\square}{\cong} (\underline{\bigodot}_{j\in[l]_{\text{odd}}} \mathbb{P}_\square^{\mathsf{X}_j|\mathsf{X}_{[j-1]}}) \tag{44}$$

$$\implies \mathbb{K} \overset{\mathbb{P}_\square}{\cong} \mathbb{P}_\square^{\mathsf{X}_l|\mathsf{X}_{[l-1]}} \tag{45}$$

Proof. The argument is that $\mathbb{K}$ and $\mathbb{P}_\square^{\mathsf{X}_l|\mathsf{X}_{[l-1]}}$ must agree except on a collection of measure 0 sets, and because the sets the union of these sets is also measure 0. See Appendix 4.6. □

If disintegrations exist, $\mathbb{K}$ is a disintegration of $\mathbb{P}_\square^{\mathsf{X}_{[n]_{\text{odd}}} \wr \mathsf{X}_{[n]_{\text{even}}}}$. Thus we can get $\mathbb{P}_\square^{\mathsf{X}_{[n]_{\text{odd}}} \wr \mathsf{X}_{[n]_{\text{even}}}}$ from $\{\mathbb{P}_\square^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in \mathsf{X}_{[n]_{\text{odd}}}\}$ by assembling and go in the reverse direction by disintegrating.

$$\begin{tikzcd}
\{\mathbb{P}_\square^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in \mathsf{X}_{[n]_{\text{odd}}}\} \arrow[rr, bend left, "\text{assemble}"] & & \mathbb{P}_\square^{\mathsf{X}_{[n]_{\text{odd}}} \wr \mathsf{X}_{[n]_{\text{even}}}} \arrow[ll, bend left, "\text{disintegrate}"]
\end{tikzcd}$$

Definition 2.33 (choices)**.** Given a probability comb $\{\mathbb{P}_\square^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in \mathsf{X}_{D_{\text{odd}}}\}$, a *choice* is a probabilty comb $\{\alpha^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in \mathsf{X}_{D_{\text{even}} \setminus \{0\}}\}$.

Probability combs define a map from choices to probability sets.

Definition 2.34 (map defined by a probability comb)**.** Given a probability comb $\{\mathbb{P}_\square^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in \mathsf{X}_{[n]_{\text{odd}}}\}$ and the set of associated choices A, we define a map $m : A \to \mathcal{P}(\Delta(\Omega))$ by

$$\{\gamma^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in \mathsf{X}_{[n]_{\text{even}}}\} \mapsto \bigodot_{i\in[n]} \mathbb{P}_\square^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} \gamma^{\mathsf{X}_{i+i}|\mathsf{X}_{[i]}} \tag{46}$$

$$:= \mathbb{P}_\gamma^{\mathsf{X}_{[n]}|\mathsf{X}_0} \tag{47}$$

In the case of an infinite index set $\mathbb{N}$, given $x_0 \in X_0$ and $\mathbb{P}_\gamma^{\mathsf{X}_{[n]}|\mathsf{X}_0}$ for all $n \in \mathbb{N}$, the Kolmogorov representation theorem guarantees that there is a unique probability $\mathbb{P}_{\gamma,x_0}^{\mathsf{X}_\mathbb{N}} \in \Delta(X^\mathbb{N})$ such that for all $n \in \mathbb{N}$, $A \in \mathcal{X}_{[n]}$

$$\mathbb{P}_{\gamma,x_0}^{\mathsf{X}_{[n]}} = \mathbb{P}_\gamma^{\mathsf{X}_{[n]}|\mathsf{X}_0}(A|x_0) \tag{48}$$

If X_0 is discrete, the map $x_0 \mapsto \mathbb{P}_{\gamma,x_0}^{\mathsf{X}_{[n]}}$ is a Markov kernel, and we define it to be $\mathbb{P}_\gamma^{\mathsf{X}_\mathbb{N}|\mathsf{X}_0}$.

Blind choices are choices that "don't depend on anything".

Definition 2.35 (blind choices)**.** Given a probability comb $\{\mathbb{P}_\square^{\mathsf{X}_i|\mathsf{X}_{[i-1]}}|i \in \mathsf{X}_{D_{\text{odd}}}\}$, a blind choice is a probabilty comb $\{\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{X}_{[i-1]}}|i \in \mathsf{X}_{D_{\text{even}}\setminus\{0\}}\}$ such that for all $i \in D_{\text{even}}$ there exists $\mathbb{K}_i : X_{[i-1]_{\text{even}}} \rightarrow X_i$ such that, up to swap isomorphism,

$$\alpha^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} = \text{erase}_{X_{[i-1]_{\text{odd}}}} \otimes \mathbb{K}_i \tag{49}$$

Theorem 2.32 is useful because for some choices α, the assembled comb is a version of the conditional probability

$$\mathbb{P}_\square^{\mathsf{X}_{[n]_{\text{odd}}} \wr \mathsf{X}_{[n]_{\text{even}}}} \cong \mathbb{P}_\alpha = \mathbb{P}_\alpha^{\mathsf{X}_{[n]_{\text{odd}}}|\mathsf{X}_{[n]_{\text{even}}}} \tag{50}$$

Note that this is almost sure with respect to $\mathbb{P}_\alpha$, not with respect to $\mathbb{P}_\square$. However, if $\mathbb{P}_\alpha^{\mathsf{X}_{[n]_{\text{odd}}}|\mathsf{X}_{[n]_{\text{even}}}}$ is unique, then it must be equal to $\mathbb{P}_\square^{\mathsf{X}_{[n]_{\text{odd}}} \wr \mathsf{X}_{[n]_{\text{even}}}}$.

Theorem 2.36 (Comb-conditional correspondence)**.** *Given a probability comb* $\{\mathbb{P}_\square^{\mathsf{X}_i|\mathsf{X}_{[i-1]}}|i \in \mathsf{X}_{D_{odd}}\}$ *and a blind choice* α

$$\mathbb{P}_\square^{\mathsf{X}_{D_{odd}} \wr \mathsf{X}_{D_{even}}} \cong \mathbb{P}_\alpha = \mathbb{P}_\alpha^{\mathsf{X}_{D_{odd}}|\mathsf{X}_{D_{even}}} \tag{51}$$

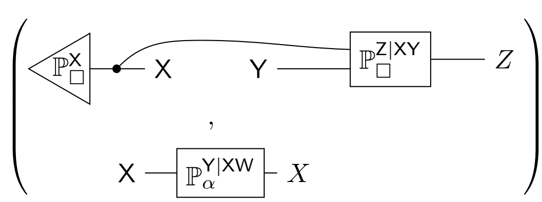
Proof. See Appendix 4.7 ∎

2.8.1 Examples of probability combs

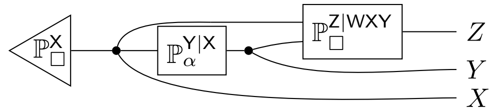
History-based reinforcement as described in Hutter (2004) is an example of a probability comb model. Such models posit sequences of variables $(\mathsf{A}_i, \mathsf{O}_i, \mathsf{R}_i)_{i\in\mathbb{N}}$ representing the ith action, observation and reward respectively. They also posit an *agent* that implements a *policy*, which is a collection of Markov kernels $\pi_i : O^i \times R^i \times A^i \rightarrow A$ 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 Markov kernels $e_i : O^i \times R^i \times A^{i+1} \rightarrow O \times R$, mapping the history of actions, rewards, observations and the next action to the next observation and reward.

Identify the envirnoment e_i with a probability comb $\{\mathbb{P}_\square^{\mathsf{O}_i\mathsf{R}_i|\mathsf{O}_{<i}\mathsf{R}_{<i}\mathsf{A}_i}|i \in \mathbb{N}\}$ and the policy set with the choices for this comb $\{\mathbb{P}_\alpha^{\mathsf{A}_i|\mathsf{O}_{<i}\mathsf{R}_{<i}\mathsf{A}_{<i}}|i \in \mathbb{N}, \alpha \in A\}$.

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 conditonal gap model graphically in an informal way as "inserting" $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}$ into $\mathbb{P}^{\mathsf{X}\square\mathsf{Z}|\mathsf{Y}}$:

Insert 

(52)

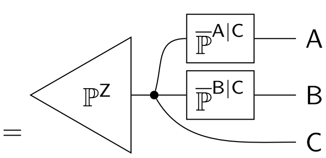
= 

(53)

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 resepect 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), see Appendix 4.3.1.

Definition 2.37 (Conditional independence with respect to a probability set)**.** For a *probability set* $\mathbb{P}_{\{\}}$ and variables $\mathsf{A}, \mathsf{B}, \mathsf{Z}$, we say B is conditionally independent of A given C, written $\mathsf{B} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{A}|\mathsf{C}$, if

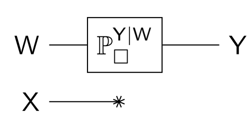
$\mathbb{P}^{\mathsf{ABC}}_{\{\}} =$ 

(54)

Cho and Jacobs (2019) have shown that this definition coincides with the standard notion of conditional independence for a particular probability model. In particular, it satisfies the *semi-graphoid axioms.*

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

Theorem 2.38. *Given standard measurable Ω, a probability model $\mathbb{P}$ and variables $\mathsf{W} : \Omega \to W$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}} \mathsf{X}|\mathsf{W}$ if and only if there exists some version of $\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}$ and $\mathbb{P}^{\mathsf{Y}|\mathsf{W}}$ such that*

$\mathbb{P}^{\mathsf{Y}|\mathsf{WX}} =$ 

(55)

$$\iff \mathbb{P}^{\mathsf{Y}|\mathsf{WX}}(y|w, x) = \mathbb{P}^{\mathsf{Y}|\mathsf{W}}(y|w) \tag{56}$$

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

The semi-graphoid axioms hold for all probability measures $\mathbb{P}$, so in particualr they hold for all $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$. Thus conditional independence with respect to a probability set also satisfies the semi-graphoid axioms.

Definition 2.39 (Conditional independence with respect to a probability comb)**.** Conditional independence $\mathsf{A} \perp\!\!\!\perp_{\mathbb{P}_\square} \mathsf{B}|\mathsf{C}$ holds for an arbitrary probability comb $\mathbb{P}_\square : A \to \mathcal{P}(\Delta(\Omega))$ if $\mathsf{A} \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{B}|\mathsf{C}$ holds for all probability models $\mathbb{P}_\alpha$, $\alpha \in A$.

2.9 Curried Markov kernels

Given a function $f : X \times Y \to Z$, we can obtain a curried version $\lambda f : Y \to Z^X$. In particular, if $Y = \{*\}$ then $\lambda f : \{*\} \to Y^X$. At least for countable X, we can apply this construction to Markov kernels: given a kernel $\mathbb{K} : X \rightarrowtriangle Y$, define $\lambda\mathbb{K} : \{*\} \rightarrowtriangle Y^X$ by

$$\lambda\mathbb{K}((y_i)_{i\in X}) = \prod_{i\in X} \mathbb{K}(y_i|i) \tag{57}$$

We can then define an evaluation map $\mathrm{ev} : Y^X \times X \to Y$ by $\mathrm{ev}((y_i)_{i\in X}, x) = y_x$. Then

$$\mathbb{K} = (\lambda\mathbb{K} \otimes \mathrm{id}_X)\mathbb{F}_{\mathrm{ev}} \tag{58}$$

Evaulation of a curried Markov kernel $\lambda\mathbb{K}$ resembles the definition of *potential outcomes*; for outcomes $\mathsf{Y} : \Omega \to Y$ and treatments $X : \Omega \to X$, potential outcomes are described by a probability distribution $\mathbb{P}^{\mathsf{Y}^X}$ on Y^X and we have the relation

$$\mathsf{Y} \overset{a.s.}{=} \mathrm{ev}(\mathsf{Y}^X, \mathsf{X}) \tag{59}$$

Then

$$\mathbb{P}^{\mathsf{Y}^X}\mathbb{F}_{\mathrm{ev}} \tag{60}$$

Is some Markov kernel $X \rightarrowtriangle Y$, which is equal to $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ if $\mathsf{Y}^X \perp\!\!\!\perp |\mathsf{X}$ which we will call the "outcome kernel". The random variable Y^X is often regarded as more than just a curried representation of outcomes, it is taken to express something like "the outcomes that would have been observed, had X turned out differently". Some authors have argued it is advantageous not to consider such interpretations unless they can be equivalently stated as assumptions about observable phenomena Richardson and Robins (2013); Dawid (2000), an approach we will stick to here.

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 $(\mathsf{X}, \mathsf{D}, \mathsf{Y})$ tell us everything we want to know for the purposes of deciding which outcomes are better than others.

We want a model that allows us to compare different stochastic *decision functions* $\mathbb{Q}_\alpha^{\mathsf{D}|\mathsf{X}} : X \rightarrowtriangle D$, 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 $\mathbb{Q}_\alpha^{\mathsf{X}|\mathsf{D}}$ and returns a probabilistic model of the consequences of selecting that decision function $\mathbb{P}_\alpha^{\mathsf{DXY}}$. An order 2 model $(\mathbb{P}_\square^{\mathsf{X}, \mathbb{P}_\square^{\mathsf{Y}}|\mathsf{XD}}, A)$ 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 $\mathsf{X}|\mathsf{D}$ given a decision function $\mathbb{Q}_\alpha^{\mathsf{X}|\mathsf{D}}$ must actually be equal to $\mathbb{Q}_\alpha^{\mathsf{X}|\mathsf{D}}$, and we can say the same for $\mathbb{P}_\square^{\mathsf{X}}$ and $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{XD}}$. 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 $(\mathbb{P}_{\square}^{\mathsf{X},\mathbb{P}_{\square}^{\mathsf{Y}}|\mathsf{XD}}, A)$ we "ought" to use. To represent this gap we include an unobserved variable H, the *hypothesis*. We can interpret H as expressing the fac that, if we knew the value of H then we would know that our decision problem was represented by a unique order 2 model $(\mathbb{P}_{h,\square}^{\mathsf{X},\mathbb{P}_{h,\square}^{\mathsf{Y}}|\mathsf{XD}}, A)$. However, H is not known and in fact we do not know how to determine H (this is the nature of an *unoserved* variable – there is no process available to find the value it yields). Our model is thus given by

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

.

Definition 3.1 (Ordinary decision problem)**.** An ordinary decision problem $(\mathbb{P}, \Omega, \mathsf{H}, (\mathsf{X}, \mathcal{X}), (\mathsf{D}, \mathcal{D}), (\mathsf{Y}, \mathcal{Y}))$ consists of a fundamental probability set Ω, hypotheses $\mathsf{H} : \Omega \to H$, observations $\mathsf{X}\Omega \to X$, decisions $\mathsf{D} : \Omega \to D$ and consequences $\mathsf{Y} : \Omega \to Y$, and the latter three random variables are associated with measurement processes. It is equipped with a probability gap model $\mathbb{P} : \Delta(D)^X \to \Delta(\Omega)^H$ where $\Delta(D)^X$ is the set of valid $\mathsf{D}|\mathsf{X}$ 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. $\mathbb{P}_\alpha^{\mathsf{D}|\mathsf{X}} = \mathbb{Q}_\alpha^{\mathsf{D}|\mathsf{X}}$ for all decision functions $\mathbb{Q}_\alpha^{\mathsf{D}|\mathsf{X}} \in \Delta(D)^X$
2. $\mathbb{P}^{\mathsf{X}|\mathsf{H}} = \mathbb{P}_\alpha^{\mathsf{X}|\mathsf{H}}$ for all $\mathbb{P}_\alpha := \mathbb{P}(\mathbb{Q}_\alpha^{\mathsf{D}|\mathsf{X}})$
3. $\mathbb{P}^{\mathsf{Y}|\mathsf{XDH}} = \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{XDH}}$ for all $\mathbb{P}_\alpha := \mathbb{P}(\mathbb{Q}_\alpha^{\mathsf{D}|\mathsf{X}})$

(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 α.

Under these assumptions $\mathbb{P}_\square$ is an order 2 model $(\mathbb{P}_{\square}^{\mathsf{X},\mathbb{P}_{\square}^{\mathsf{Y}}|\mathsf{XD}}, A)$ 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 measurment 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 $X \times D \times Y \to \mathbb{R}$ 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 "decisons" share and this distinction might be important regarding judgements of causal contractibilty.

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

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

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

A *statistical model* (or *statistical experiment*) is a collection of probability distributions $\{\mathbb{P}_\theta\}$ indexed by some set Θ. A statistical decision problem gives us an observation variable $\mathsf{X} : \Omega \rightarrow X$ and a statistical experiment $\{\mathbb{P}^{\mathsf{X}}_\theta\}_\Theta$, a decision set D and a loss $l : \Theta \times D \to \mathbb{R}$. A strategy $\mathbb{S}^{\mathsf{D}|\mathsf{X}}_\alpha$ is evaluated according to the risk functional $R(\theta, \alpha) := \sum_{x\in X} \sum_{d\in D} \mathbb{P}^{\mathsf{X}}_\theta(x) S^{\mathsf{D}|\mathsf{X}}_\alpha(d|x) l(h, d)$. A strategy $\mathbb{S}^{\mathsf{D}|\mathsf{X}}_\alpha$ is considered more desirable than $\mathbb{S}^{\mathsf{D}|\mathsf{X}}_\beta$ if $R(\theta, \alpha) < R(\theta, \beta)$.

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

$$\mathbb{E}_{\mathbb{P}_\alpha}[\mathsf{Y}|h] = \sum_{x\in X} \sum_{d\in D} \sum_{y\in Y} \mathbb{P}^{\mathsf{X}|\mathsf{H}}(x|h)\mathbb{Q}^{\mathsf{D}|\mathsf{X}}_\alpha(d|x)\mathbb{P}^{\mathsf{Y}|\mathsf{HD}}(y|h, d) \tag{61}$$

$$= \sum_{x\in X} \sum_{d\in D} \mathbb{P}^{\mathsf{X}|\mathsf{H}}(x|h)\mathbb{Q}^{\mathsf{D}|\mathsf{X}}_\alpha(d|x) l(h, d) \tag{62}$$

$$= R(h, \alpha) \tag{63}$$

If we are given a see-do model where we interpret $\{\mathbb{P}^{\mathsf{X}|\mathsf{H}}(\cdot|h)|h \in H\}$ as a statistical experiment and Y as a negative utility, the expectation of the utility under the strategy forecast given in equation **??** 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 S appropriate for modelling $\mathcal{S}$ and

    1. The same model is appropriate for any measurement process that first peforms $\mathcal{S}$ and subsequently shuffles the results according to any permutation $\text{swap}_a : S^n \to S^n$ 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 $(\mathbb{P}, S^{|A|}, \mathcal{S}^{|A|})$ is exchangeable, I mean

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

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 $\mathbb{P}^{\mathsf{S}}$ 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 indpendent 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 at first consider *blind* decision functions – that is, decision functions that pay no attention to the data already available. We are interested in repetitive symmetries, and these symmetries are typically broken if our decisions are based on past observations. Furthermore, we can define the entire see-do model by its behaviour on blind decision functions. We also assume that the hypotheses are trivial $\mathsf{H} = *$; 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 sqeuences 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 decisions 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 contractibly 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 go on to prove that contractible causal models can also be represented by jointly independent repetitions of a "unit-level consequence map", indexed by a hypothesis H. By currying this consequence map, we can also obtain a "potential outcomes-like" representation of consequences.

A probability comb "commutes with exchange" if applying any finite permutation to blind choices or separately applying the corresponding permuation to consequences each yields the same result. The term *commute* comes from the notion that we can apply the exchange "before" the conditional $\mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{D}}$ or after it and get the same result.

Definition 4.1 (Commutativity of exchange)**.** Suppose we have a fundamental probability set Ω and a do model $(\mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{D}}, R)$ such that $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$. For a finite permutation $\rho : \mathbb{N} \to \mathbb{N}$, define $\mathrm{swap}_{\rho(D)} : D \rightarrow D$ by

$(d_i)_{i\in\mathbb{N}} \mapsto \delta_{(d_{\rho(i)})_{i\in\mathbb{N}}}$ and $\text{swap}_{\rho(D\times Y)} : D \times Y \rightharpoonup D \times Y$ by $(x_i)_{i\in\mathbb{N}} \mapsto \delta_{(x_{\rho(i)})_{i\in\mathbb{N}}}$. If, for any two decision rules $\alpha, \beta \in R$,

$$\mathbb{P}_\alpha^{\mathsf{D}} \text{swap}_{\rho(D)} = \mathbb{P}_\beta^{\mathsf{D}} \tag{64}$$

$$\implies \mathbb{P}_\alpha \text{swap}_{\rho(D\times Y)} = \mathbb{P}_\beta \tag{65}$$

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 $\mathsf{D} = (\mathsf{D}_1, \mathsf{D}_2, \mathsf{D}_3)$ and $\mathsf{Y} = (\mathsf{Y}_1, \mathsf{Y}_2, \mathsf{Y}_3)$ and $\mathbb{P}_\alpha^{\mathsf{D}_1\mathsf{D}_3} = \mathbb{P}_\beta^{\mathsf{D}_3\mathsf{D}_2}$ then $\mathbb{P}_\alpha^{\mathsf{Y}_1\mathsf{Y}_3} = \mathbb{P}_\beta^{\mathsf{Y}_3\mathsf{Y}_2}$.

Definition 4.2 (Causal contractibility)**.** Suppose we have a fundamental probability set Ω and a do model $(\mathbb{P}, \mathsf{D}, \mathsf{Y}, R)$ such that $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$. 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 $\mathsf{D}_S := (\mathsf{D}_i)_{i\in S}$ and $\mathsf{D}_T := (\mathsf{D}_i)_{i\in T}$. If for any $\alpha, \beta \in R$

$$\mathbb{P}_\alpha^{\mathsf{D}_S} = \mathbb{P}_\beta^{\mathsf{D}_T} \implies \mathbb{P}_\alpha^{(\mathsf{D_i},\mathsf{Y_i})_{i\in S}} = \mathbb{P}_\beta^{(\mathsf{D_i},\mathsf{Y_i})_{i\in T}} \tag{66}$$

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

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 $\alpha \in R$

$$\mathbb{P}_\alpha^{\mathsf{Y}_1\mathsf{Y}_2|\mathsf{D}_1\mathsf{D}_2}(y_1, y_2|d_1, d_2) = [\![(y_1, y_2) = (d_1 + d_2, d_1 + d_2)]\!] \tag{67}$$

Then $\mathbb{P}_{00}^{\mathsf{Y}_1}(y_1) = [\![y_1 = 0]\!]$ while $\mathbb{P}_{01}^{\mathsf{Y}_1} = [\![y_1 = 1]\!]$, 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),

$$\mathbb{P}_{d_2,d_1}^{\mathsf{Y}_1\mathsf{Y}_2|\mathsf{D}_1\mathsf{D}_2}(y_1, y_2) = [\![(y_1, y_2) = (d_2 + d_1, d_2 + d_1)]\!] \tag{68}$$

$$= [\![(y_2, y_1) = (d_1 + d_2, d_1 + d_2)]\!] \tag{69}$$

$$= \mathbb{P}_{d_1,d_2}^{\mathsf{Y}_1\mathsf{Y}_2|\mathsf{D}_1\mathsf{D}_2}(y_2, y_1) \tag{70}$$

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, $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}} (\mathsf{D}_j, \mathsf{Y}_j)_{j\in\mathbb{N}} \setminus \{i\}|\mathsf{H}\mathsf{D}_i$, where H is a symmetric function of $(\mathsf{Y}_i, \mathsf{D}_i)_{i\in\mathbb{N}}$.

4.2 Representations of contractible probability models

We prove two representation theorems for causally contractible do models. Theorem 4.4 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" matrix 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. However, as we will point out, this lookup table may or may not support an interpretation as a table of potential outcomes. Furthermore, we make use of this theorem in proving Theorem 4.6, 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.

Definition 4.3 (Contractible probability distribution)**.** Given a fundamental probability set Ω, variable $\mathsf{X} := (\mathsf{X}_i)_{i\in\mathbb{N}}$ and a probability distribution $\mathbb{P}^{\mathsf{X}} \in \Delta(X^{\mathbb{N}})$, 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 $\mathsf{X}_S := (\mathsf{X}_i)_{i\in S}$ and $\mathsf{X}_T := (\mathsf{X}_i)_{i\in T}$. If

$$\mathbb{P}^{\mathsf{X}_S} = \mathbb{P}^{\mathsf{X}_T} \tag{71}$$

$\mathbb{P}$ 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 $\mathsf{Y}^D : \Omega \to Y^{D\times\mathbb{N}}$ which can be represented as a matrix of variables Y_{ij}

$$\mathsf{Y}^D = \begin{array}{cc} & \overset{\mathbb{N}\text{ columns}}{\longleftrightarrow} \\ |D| \text{ rows} \updownarrow & \begin{matrix} \mathsf{Y}_{11} & \mathsf{Y}_{12} & \mathsf{Y}_{13} & \mathsf{Y}_{14} & \\ \mathsf{Y}_{21} & \mathsf{Y}_{22} & \mathsf{Y}_{23} & \mathsf{Y}_{24} & \cdots \\ \mathsf{Y}_{31} & \mathsf{Y}_{32} & \mathsf{Y}_{33} & \mathsf{Y}_{34} & \end{matrix} \end{array} \tag{72}$$

and, given any deterministic decision function δ_d, $d = (d_i)_{i\in\mathbb{N}} \in D^{\mathbb{N}}$, we can find $\mathbb{P}^{\mathsf{Y}|\mathsf{D}}$ by "looking up" d in the table. For example, if $d = (1, 2, 3, 2, ...)$, Equation 73 illustrates the idea of "looking up" the relevant elements of Y^D and Equation 74 illustrates the resulting value of $\mathbb{P}^{\mathsf{Y}|\mathsf{D}}$.

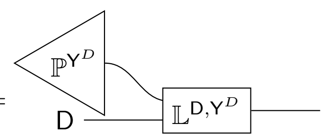
$$\begin{array}{rccccc} d = & 1 & 2 & 3 & 2 & \cdots \\ & \mathsf{Y}_{11} & \mathsf{Y}_{12} & \mathsf{Y}_{13} & \mathsf{Y}_{14} & \\ \mathsf{Y}^D = & \mathsf{Y}_{21} & \mathsf{Y}_{22} & \mathsf{Y}_{23} & \mathsf{Y}_{24} & \cdots \\ & \mathsf{Y}_{31} & \mathsf{Y}_{32} & \mathsf{Y}_{33} & \mathsf{Y}_{34} & \end{array} \tag{73}$$

$$\mathbb{P}^{\mathsf{Y}|\mathsf{D}}(y|(1, 2, 3, 2, ...)) = \mathbb{P}^{\mathsf{Y}_1 1\mathsf{Y}_2 2\mathsf{Y}_{33}\mathsf{Y}_{24}\cdots}(y) \tag{74}$$

The contractibility of $\mathbb{P}^{\mathsf{Y}^D}$ means that any two subcollections of columns of the same size are equal in distribution, and the exchangeability of $\mathbb{P}^{\mathsf{Y}^D}$ means that the random variable obtained by permuting its columns is also equal in distribution to 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 4.4 (Table representation of causally contractible do models)**.** *Suppose we have a fundamental probability set Ω and a do model $(\mathbb{P}, \mathsf{D}, \mathsf{Y}, R)$ such that $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$. $\mathbb{P}$ is causally contractible if and only if*

$\mathbb{P}^{\mathsf{Y}|\mathsf{D}} =$  (75)

$$\iff \tag{76}$$

$$\mathbb{P}^{\mathsf{Y}|\mathsf{D}}(y|d) = \mathbb{P}^{(\mathsf{Y}^D_{d_i i})_{\mathbb{N}}}(y) \tag{77}$$

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

$$l : D^{\mathbb{N}} \times Y^{D\times\mathbb{N}} \to Y \tag{78}$$

$$((d_i)_{\mathbb{N}}, (y_{ij})_{i\in D, j\in\mathbb{N}}) \mapsto y_{d_i i} \tag{79}$$

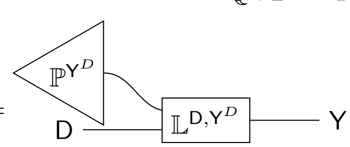
Proof. Only if: Choose $e := (e_i)_{i\in\mathbb{N}}$ such that $e_{|D|i+j}$ is the ith element of D for all $i, j \in \mathbb{N}$. Abusing notation, write e also for the decision function that chooses e deterministically.

Define

$$\mathbb{P}^{\mathsf{Y}^D}((y_{ij})_{D\times\mathbb{N}}) := \mathbb{P}^{\mathsf{Y}}_e((y_{|D|i+j})_{i\in D, j\in\mathbb{N}}) \tag{80}$$

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

$$\mathbb{Q} : D \rightarrow Y \tag{81}$$

$\mathbb{Q} :=$  (82)

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

$$
\begin{aligned}
\sum_{y \in \mathsf{Y}^{-1}(y_A)} \mathbb{Q}(y|d) &= \sum_{y \in \mathsf{Y}^{-1}(y_A)} \mathbb{P}^{(\mathsf{Y}^D_{d_i i})_A}(y) && (83)\\
&= \sum_{y \in \mathsf{Y}^{-1}(y_A)} \mathbb{P}_e^{(\mathsf{Y}_{|D|d_i+i})_A}(y) && (84)\\
&= \mathbb{P}_e^{\mathsf{Y}_B}(y_A) && (85)\\
&= \mathbb{P}_d^{\mathsf{Y}_A}(y_A) \qquad \text{by causal contractibility} && (86)
\end{aligned}
$$

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

$$\mathbb{Q}(y|d) = \mathbb{P}_d^{\mathsf{Y}}(y) \tag{87}$$

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

$$\mathbb{Q}(y|d) = \mathbb{P}_d^{\mathsf{Y}|\mathsf{D}}(y|d) \tag{88}$$

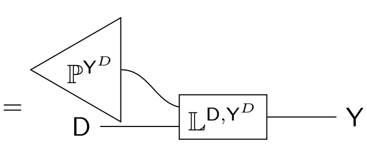
And because $\mathbb{P}_d^{\mathsf{Y}|\mathsf{D}}(y|d)$ is unique for all $d \in D^{\mathbb{N}}$ and $\mathbb{P}^{\mathsf{Y}|\mathsf{D}}$ exists by assumption

$$\mathbb{P}^{\mathsf{Y}|\mathsf{D}} = \mathbb{Q} \tag{89}$$

Next we will show $\mathbb{P}^{\mathsf{Y}^D}$ is contractible. Consider any subsequences Y^D_S and Y^D_T of 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

$$
\begin{aligned}
\mathbb{P}^{\mathsf{Y}^D_S} &= \mathbb{P}_e^{\mathsf{Y}_{\rho(S)}}) && (90)\\
&= \mathbb{P}_e^{\mathsf{Y}_{\rho(T)}}) \quad \text{by contractibility of } \mathbb{P} \text{ and the equality } e_{\rho(S)} = e_{\rho(T)} && (91)\\
&= \mathbb{P}^{\mathsf{Y}^D_T} && (92)
\end{aligned}
$$

If: Suppose

$\mathbb{P}^{\mathsf{Y}|\mathsf{D}} =$  (93)

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

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

By definition,

$$\mathbb{P}^{\mathsf{Y}_S|\mathsf{D}}(y_S|d) = \sum_{y_S^D \in Y^{|D|\times|S|}} \mathbb{P}^{\mathsf{Y}_S^D}(y_S^D)\mathbb{L}^{\mathsf{D}_S,\mathsf{Y}^S}(y_S|d, y_S^D) \tag{94}$$

$$= \sum_{y_S^D \in Y^{|D|\times|T|}} \mathbb{P}^{\mathsf{Y}_T^D}(y_S^D)\mathbb{L}^{\mathsf{D}_S,\mathsf{Y}^S}(y_S|d, y_S^D) \quad \text{by contractibility of } \mathbb{P}^{\mathsf{Y}_T^D} \tag{95}$$

$$= \mathbb{P}^{\mathsf{Y}_T|\mathsf{D}}(y_S|d) \tag{96}$$

□

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. Our argument for a potential outcomes representation does not go through in this case, because it hinges on the fact that we can "wrap" the outcomes under a particular blind decision into a table, and then use contractibility to choose one outcome from each column, however using contractibility also gives us exchangeability of the columns.

It is also worth noting that the lookup table does not need to have an interpretation as a collection of potential outcomes. For example, consider a series of bets on fair coinflips – in this case, the consequence Y_i is uniform on $\{0,1\}$ for any decision D_i. Tha $D = Y = \{0,1\}$ and $\mathbb{P}_\alpha^{\mathsf{Y}_n}(y) = \prod_{i\in[n]} 0.5$ for all n, $y \in Y^n$, $\alpha \in R$. Then the construction in Theorem 4.4 yields $\mathbb{P}^{Y_i^D}(y_i^D) = \prod_{j\in D} 0.5$ for all $y_i^D \in Y^D$. That is, Y_i^0 and Y_i^1 are independent and uniformly distributed. However, if we wanted Y_i^0 to represent "what would happen if I bet 0 on turn i" and Y^1 to represent "what would happen if I bet 1 on turn i", then we actually want $\mathsf{Y}_i^0 = 1 - \mathsf{Y}_i^1$. Thus the measurement table lookup is formally similar to the potential outcomes setup, but potential outcomes attributes additional semantics to the entries in the lookup table which can impose extra requirements on their distribution.

Theorem 4.5 establishes a claim made earlier: that contractibility is strictly stronger than commutativity of exchange.

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

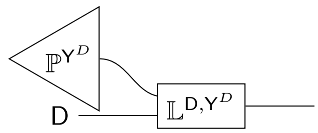
Proof. Given a finite permutation $\rho : \mathbb{N} \to \mathbb{N}$ and any sequence $x := (x_i)_{i\in\mathbb{N}}$ let $\rho(x) = (x_{\rho(i)})_{i\in\mathbb{N}}$ or equivalently $(x_i)_{i\in\rho(\mathbb{N})}$. Then for any $d = (d_i)_{i\in\mathbb{N}}$ and $y^D := (y_{ij})_{i\in D, j\in\mathbb{N}}$:

$$l(\rho(d), y^D) = (y_{d_{\rho(i)} i})_{i\in\mathbb{N}} \tag{97}$$

$$= (y_{d_i \rho^{-1}(i)})_{i\in\rho(\mathbb{N})} \tag{98}$$

$$= \rho(l(d, \rho^{-1}(y^D))) \tag{99}$$

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

$\mathbb{P}^{\mathsf{Y}|\mathsf{D}} =$  (100)

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

$$\mathbb{P}^{\mathsf{Y}}_{\rho(d)}(y) = \sum_{y'^D \in Y^{D\times\mathbb{N}}} [\![l_{DY}(\rho(d), y'^D) = y]\!] \mathbb{P}^{\mathsf{Y}^D}(y'^D) \tag{101}$$

$$= \sum_{y'^D \in Y^{D\times\mathbb{N}}} [\![l_{DY}(d, \rho^{-1}(y'^D)) = \rho^{-1}(y)]\!] \mathbb{P}^{\mathsf{Y}^D}(y'^D) \tag{102}$$

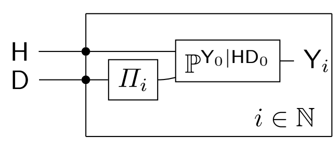
$$= \sum_{y'^D \in Y^{D\times\mathbb{N}}} [\![l_{DY}(d, \rho^{-1}(y'^D)) = \rho^{-1}(y)]\!] \mathbb{P}^{\mathsf{Y}^D}(\rho^{-1}(y'^D)) \tag{103}$$

$$= \mathbb{P}^{\mathsf{Y}}_{\rho(d)}(\rho^{-1}(y)) \tag{104}$$

□

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 4.6. *Suppose we have a fundamental probability set Ω and a do model $(\mathbb{P}, \mathsf{D}, \mathsf{Y}, R)$ such that $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$. $\mathbb{P}$ is causally contractible if and only if there exists some $\mathsf{H} : \Omega \to H$ such that $\mathbb{P}^{\mathsf{Y}_i|\mathsf{HD}_i}$ exists for all $i \in \mathbb{N}$ and*

$\mathbb{P}^{\mathsf{Y}|\mathsf{HD}} =$  (105)

$$\iff \tag{106}$$

$$\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}} \mathsf{Y}_{\mathbb{N}\setminus i}, \mathsf{D}_{\mathbb{N}\setminus i} | \mathsf{HD}_i \qquad \forall i \in \mathbb{N} \tag{107}$$

$$\wedge \mathbb{P}^{\mathsf{Y}_i|\mathsf{HD}_i} = \mathbb{P}^{\mathsf{Y}_0|\mathsf{HD}_0} \qquad \forall i \in \mathbb{N} \tag{108}$$

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

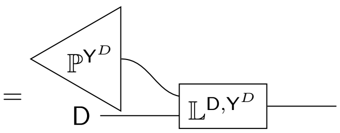
$$\mathbb{P}^{\mathsf{Y}_S|\mathsf{HD}}_d(y|d) = \int_H \prod_{i\in S} \mathbb{P}^{\mathsf{Y}_0|\mathsf{HD}_0}(y_i|h, d_i) d\mathbb{P}^{\mathsf{H}}(h) \tag{109}$$

$$= \int_H \prod_{i\in T} \mathbb{P}^{\mathsf{Y}_0|\mathsf{HD}_0}(y_i|h, d'_i) d\mathbb{P}^{\mathsf{H}}(h) \quad \text{by equality of subsequences} \tag{110}$$

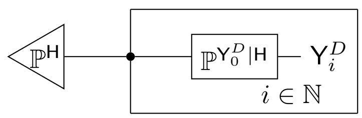
$$= \mathbb{P}^{\mathsf{Y}_T|\mathsf{HD}}_{d'}(y|d) \tag{111}$$

Only if: We have

$$\mathbb{P}^{\mathsf{Y}|\mathsf{D}} = \tag{112}$$



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



$$\mathbb{P}^{\mathsf{Y}^{\mathsf{D}}} = \tag{113}$$

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

$$\begin{aligned}
&\mathsf{Y}_{\cdot i}^D \perp\!\!\!\perp_{\mathbb{P}} \mathsf{Y}_{\cdot\{i\}^C}^D|\mathsf{H} && \text{representation theorem} && (114)\\
&\mathsf{Y}^D\mathsf{H} \perp\!\!\!\perp_{\mathbb{P}} \mathsf{D} && \text{by Theorem 4.11 and existence of } \mathbb{P}^{\mathsf{Y}^D\mathsf{H}} && (115)\\
&\mathsf{Y}_{\cdot i}^D \perp\!\!\!\perp_{\mathbb{P}} \mathsf{D}|\mathsf{Y}_{\cdot\{i\}^C}^D\mathsf{H} && \text{weak union on Eq. 115} && (116)\\
&\mathsf{Y}_{\cdot i}^D \perp\!\!\!\perp_{\mathbb{P}} \mathsf{D}\mathsf{Y}_{\cdot\{i\}^C}^D|\mathsf{H} && \text{contraction on Eqs. 114 and 115} && (117)\\
&\mathsf{Y}_{\cdot i}^D \perp\!\!\!\perp_{\mathbb{P}} \mathsf{D}_{\{i\}^C}\mathsf{Y}_{\cdot\{i\}^C}^D|\mathsf{H}\mathsf{D}_i && \text{weak union on Eq. 117} && (118)\\
&\mathsf{D}_i \perp\!\!\!\perp_{\mathbb{P}} \mathsf{Y}_{\cdot\{i\}^C}^D\mathsf{D}_{\{i\}^C}|\mathsf{H}\mathsf{D}_i\mathsf{Y}_{\cdot i}^D && \text{due to conditioning on } \mathsf{D}_i && (119)\\
&\mathsf{Y}_i^D\mathsf{D}_i \perp\!\!\!\perp_{\mathbb{P}} \mathsf{D}_{\{i\}^C}\mathsf{Y}_{\cdot\{i\}^C}^D|\mathsf{H}\mathsf{D}_i && \text{contraction on Eqs. 118 and 119} && (120)\\
& && && (121)
\end{aligned}$$

Now, note that $(\mathsf{Y}_i, \mathsf{D}_i)$ is a deterministic function of $(\mathsf{Y}_i^D, \mathsf{D}_i)$ and $(\mathsf{Y}_{\{i\}^C}, \mathsf{D}_{\{i\}^C})$ is a deterministic function of $(\mathsf{Y}_{\{i\}^C}^D, \mathsf{D}_{\{i\}^C})$. Therefore

$$\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}} \mathsf{D}_{\{i\}^C}\mathsf{Y}_{\{i\}^C}|\mathsf{H}\mathsf{D}_i \tag{122}$$

So, by Theorem 4.11, $\mathbb{P}^{\mathsf{Y}_i|\mathsf{HD}}$ exists and by contractibility of $\mathbb{P}^{\mathsf{Y}^D}$, for any $i, j \in \mathbb{N}$

$$\begin{aligned}
\mathbb{P}^{\mathsf{Y}_i|\mathsf{HD}_\mathsf{i}}(y_i|h, d_i) &= \mathbb{P}^{\mathsf{Y}_{d_i i}^D|\mathsf{H}}(y_i|h) && (123)\\
&= \mathbb{P}^{\mathsf{Y}_{d_i j}^D|\mathsf{H}}(y_i|h) && (124)\\
&= \mathbb{P}^{\mathsf{Y}_j|\mathsf{HD}_j}(y_i|h, d_i) && (125)
\end{aligned}$$

□

4.3 Potential outcomes

4.3.1 Extended conditional independence

Needs a support condition

In the case of a probability gap model $(\mathbb{P}_\square^{\mathsf{V}|\mathsf{W}}, A)$ 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 4.7. *Let* $\mathsf{A}^* = \mathsf{A} \circ \mathsf{V}$*,* $\mathsf{B}^* = \mathsf{B} \circ \mathsf{V}$*,* $\mathsf{C}^* = \mathsf{C} \circ \mathsf{V}$ *(*$(\mathsf{A}, \mathsf{B}, \mathsf{C})$ *are* $\mathcal{V}$*-measurable) and* $\mathsf{D}^* = \mathsf{D} \circ \mathsf{W}$*,*$\mathsf{E}^* = \mathsf{E} \circ \mathsf{W}$ *where* W *is discrete and* $\mathsf{W} = (\mathsf{D}^*, \mathsf{E}^*)$*. In addition, let* $\mathbb{P}_\alpha^{\mathsf{W}}$ *be some probability distribution on* W *such that* $w \in \mathsf{W}(\Omega) \implies \mathbb{P}_\alpha^{\mathsf{W}}(w) > 0$*. Then, denoting extended conditional independence with* $\perp\!\!\!\perp_{\mathbb{P},ext}$ *and* $\mathbb{P}_\alpha^{\mathsf{VW}} := \mathbb{P}_\alpha^{\mathsf{W}} \odot \mathbb{P}^{\mathsf{V}|\mathsf{W}}$

$$\mathsf{A} \perp\!\!\!\perp_{\mathbb{P},ext} (\mathsf{B}, \mathsf{D})|(\mathsf{C}, \mathsf{E}) \iff \mathsf{A}^* \perp\!\!\!\perp_{\mathbb{P}_\alpha} (\mathsf{B}^*, \mathsf{D}^*)|(\mathsf{C}^*, \mathsf{E}^*) \tag{126}$$

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

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

Theorem 4.8. *Let* $\mathsf{A}^* = \mathsf{A} \circ \mathsf{V}$*,* $\mathsf{B}^* = \mathsf{B} \circ \mathsf{V}$*,* $\mathsf{C}^* = \mathsf{C} \circ \mathsf{V}$ *(*$(\mathsf{A}, \mathsf{B}, \mathsf{C})$ *are* $\mathcal{V}$*-measurable) and* $\mathsf{D}^* = \mathsf{D} \circ \mathsf{W}$*,*$\mathsf{E}^* = \mathsf{E} \circ \mathsf{W}$ *where* V, W *are discrete and* $\mathsf{W} = (\mathsf{D}^*, \mathsf{E}^*)$*. Then letting* $\mathbb{P}_\alpha^{\mathsf{VW}} := \mathbb{P}_\alpha^{\mathsf{W}} \odot \mathbb{P}^{\mathsf{V}|\mathsf{W}}$

$$\mathsf{A} \perp\!\!\!\perp^1_{\mathbb{P},ext} (\mathsf{B}, \mathsf{D})|(\mathsf{C}, \mathsf{E}) \iff \mathsf{A}^* \perp\!\!\!\perp_{\mathbb{P}} (\mathsf{B}^*, \mathsf{D}^*)|(\mathsf{C}^*, \mathsf{E}^*) \tag{127}$$

Proof. If:

By assumption, $\mathsf{A}^* \perp\!\!\!\perp_{\mathbb{P}_\alpha} (\mathsf{B}^*, \mathsf{D}^*)|(\mathsf{C}^*, \mathsf{E}^*)$ for all $\mathbb{P}_\alpha^{\mathsf{D}^*\mathsf{E}^*}$. In particular, this holds for some $\mathbb{P}_\alpha^{\mathsf{D}^*\mathsf{E}^*}$ such that $(d, e) \in (\mathsf{D}^*, \mathsf{E}^*)(\Omega) \implies \mathbb{P}_\alpha^{\mathsf{D}^*\mathsf{E}^*}(d, e) > 0$. Then by Theorem 4.7, $\mathsf{A} \perp\!\!\!\perp_{\mathbb{P},\mathrm{ext}} (\mathsf{B}, \mathsf{D})|(\mathsf{C}, \mathsf{E})$.

Only if:

For any β, $\mathbb{P}_\beta^{\mathsf{ABC}|\mathsf{DE}} = \mathbb{P}_\beta^{\mathsf{DE}} \odot \mathbb{P}^{\mathsf{ABC}|\mathsf{DE}}$. By Lemma 2.14, we have $\mathbb{P}^{\mathsf{A}|\mathsf{BCDE}}$ such that

$$\mathbb{P}_\beta^{\mathsf{A}^*\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} = \mathbb{P}_\beta^{\mathsf{D}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{B}^*\mathsf{C}^*|\mathsf{D}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{A}^*|\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} \tag{128}$$

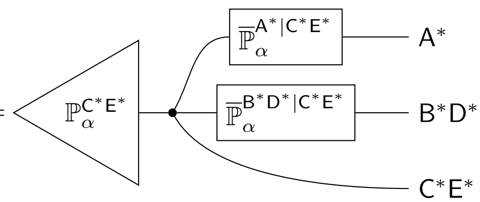
$$= \mathbb{P}_\beta^{\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{A}^*|\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} \tag{129}$$

$$= \mathbb{P}_\beta^{\mathsf{C}^*\mathsf{E}^*} \odot \mathbb{P}_\beta^{\mathsf{B}^*\mathsf{D}^*|\mathsf{C}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{A}^*|\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} \tag{130}$$

By Theorem 4.7, we have some α such that $\mathbb{P}_\alpha^{\mathsf{D}^*\mathsf{E}^*}$ is strictly positive on the range of $(\mathsf{D}^*, \mathsf{E}^*)$ and $\mathsf{A}^* \perp\!\!\!\perp_{\mathbb{P}_\alpha} (\mathsf{B}^*, \mathsf{D}^*)|(\mathsf{C}^*, \mathsf{E}^*)$.

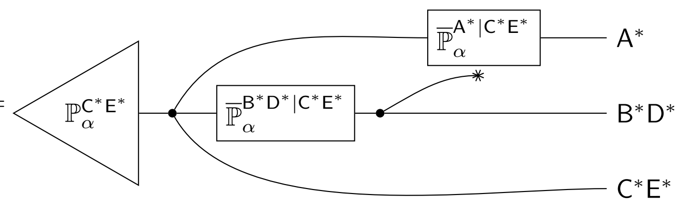
By independence, for some version of $\mathbb{P}^{\mathsf{A}|\mathsf{BCDE}}$:

$$\mathbb{P}_\alpha^{\mathsf{C}^*\mathsf{E}^*} \odot \mathbb{P}_\alpha^{\mathsf{B}^*\mathsf{D}^*|\mathsf{C}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{A}^*|\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} =$$



(131)

$$=$$

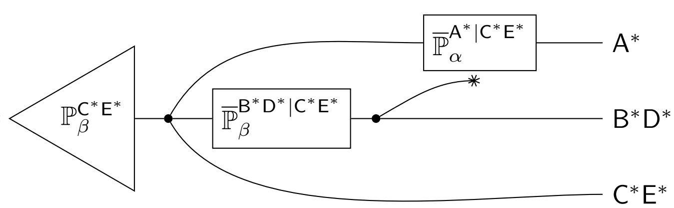


(132)

$$= \mathbb{P}_\alpha^{\mathsf{C}^*\mathsf{E}^*} \odot \mathbb{P}_\alpha^{\mathsf{B}^*\mathsf{D}^*|\mathsf{C}^*\mathsf{E}^*} \odot (\mathbb{P}_\alpha^{\mathsf{A}^*|\mathsf{C}^*\mathsf{E}^*} \otimes \text{erase}_{BD}) \tag{133}$$

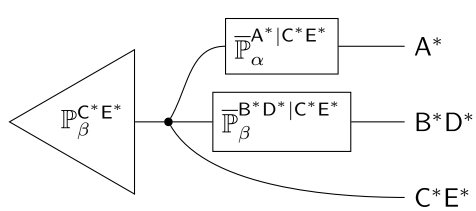
Thus for any $(a,b,c,d,e) \in A\times B\times C\times D\times E$ such that $\mathbb{P}_\alpha^{\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*}(b,c,d,e) > 0$, $\mathbb{P}^{\mathsf{A}^*|\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*}(a|b,c,d,e) = \mathbb{P}_\alpha^{\mathsf{A}^*|\mathsf{C}^*\mathsf{E}^*}(a|c,e)$.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

$$\mathbb{P}_\beta^{\mathsf{A}^*\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} =$$



(134)

$$=$$



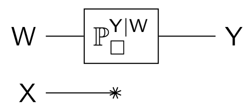
(135)

□

We can deduce conditional independences in probability combs when conditional probabilities exist and they are *unresponsive* to some input variables.

Definition 4.9 (Unresponsiveness)**.** Given discrete Ω, a probability gap model $\mathbb{P}_\square : A \to \Delta(\Omega)$, variables $\mathsf{W} : \Omega \to W$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, if there is some version of the conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}$ and $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{W}}$ such that

$$\mathbb{P}_\square^{\mathsf{Y}|\mathsf{WX}} = \tag{136}$$



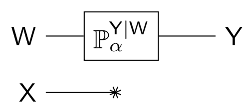
then $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{WX}}$ is *unresponsive* to X.

Definition 4.10 (Domination)**.** Given a probability gap model $\mathbb{P}_\square : A \to \Delta(\Omega)$, $\alpha \in A$ dominates A if $\mathbb{P}_\beta(B) > 0 \implies \mathbb{P}_\alpha(B) > 0$ for all $\beta in A$, $B \in \mathcal{F}$.

Theorem 4.11 (Conditional independence from kernel unresponsiveness)**.** *Given discrete Ω, variables $\mathsf{W} : \Omega \to W$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and a probability gap model $\mathbb{P}_\square : A \to \Delta(\Omega)$ with conditional probability $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{WX}}$ and such that there is $\alpha \in A$ dominating A, $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_\square} \mathsf{X}|\mathsf{W}$ if and only if $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{WX}}$ is unresponsive to W.*

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

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}} = \tag{137}$$



And so, by Theorem 2.38, $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{X}|\mathsf{W}$ for all $\alpha \in A$, and so $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_\square} \mathsf{X}|\mathsf{W}$. Only if: For α dominating A, by Theorem 2.38, there exists a version of $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}}$ unresponsive to W. Because α dominates A, every version of $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}}$ is a version of $\mathbb{P}_\beta^{\mathsf{Y}|\mathsf{WX}}$ for all $\beta in A$, thus it is a version of $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{WX}}$ also. □

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

4.3.2 Graphical properties of conditional independence

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

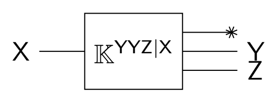
We can reason from graphical properties of model disintegrations to graphical properties of models as Theorem 4.11. A general theory akin to d-separation for string diagrams may facilitate a more general understanding of how conditional independence properties of a model relate to conditional independence properties of its components.

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

4.4.1 Repeated variables

Lemmas 4.12 and 4.13 establish that models of repeated variables must connect the repetitions with a copy map.

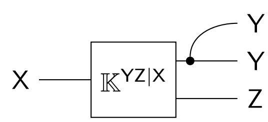
Lemma 4.12 (Output copies of the same variable are identical)**.** *For any Ω, $\mathsf{X}, \mathsf{Y}, \mathsf{Z}$ random variables on Ω and conditional probability $\mathbb{K}^{\mathsf{YZ}|\mathsf{X}}$, there is a conditional probability $\mathbb{K}^{\mathsf{YYZ}|\mathsf{X}}$ unique up to impossible values of X such that*



$$= \mathbb{K}^{\mathsf{YZ}|\mathsf{X}} \tag{138}$$

and it is given by

$$\mathbb{K}^{\mathsf{YYZ}|\mathsf{X}} = \tag{139}$$



$$\iff \tag{140}$$

$$\mathbb{K}^{\mathsf{YYZ}|\mathsf{X}}(y, y', z|x) = [\![y = y']\!]\mathbb{K}^{\mathsf{YZ}|\mathsf{X}}(y, z|x) \tag{141}$$

$$\tag{142}$$

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

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

$$\mathbb{K}^{\mathsf{YYZ}|\mathsf{X}}(y, y', z|x) = [\![y = y']\!]\mathbb{L}(y, z|x) \tag{143}$$

$$\tag{144}$$

But then

$$\mathbb{K}^{\mathsf{YZ}|\mathsf{X}}(y, z|x) = \sum_{y' \in Y} \mathbb{K}^{\mathsf{YYZ}|\mathsf{X}}(y, y', z|x) \tag{145}$$

$$= \mathbb{L}(y, z|x) \tag{146}$$

$$\tag{147}$$

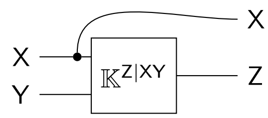
□

Lemma 4.13 (Copies shared between input and output are identical)**.**

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

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

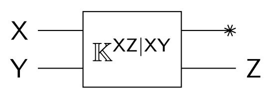
$$\mathbb{K} = \tag{148}$$



$$\iff \tag{149}$$

$$\mathbb{K}'^{x',z}_{x,y} = [\![x = x']\!]\mathbb{L}^{z}_{'x,y} \tag{150}$$

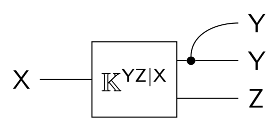
For any Ω*,* $\mathsf{X}, \mathsf{Y}, \mathsf{Z}$ *random variables on* Ω *and conditional probability* $\mathbb{K}^{\mathsf{Z}|\mathsf{XY}}$*, there is a conditional probability* $\mathbb{K}^{\mathsf{XZ}|\mathsf{XY}}$ *unique up to impossible values of* (X, Y) *such that*



$$= \mathbb{K}^{\mathsf{XZ}|\mathsf{XY}} \tag{151}$$

and it is given by

$$\mathbb{K}^{\mathsf{XZ}|\mathsf{XY}} = \tag{152}$$



$$\iff \tag{153}$$

$$\mathbb{K}^{\mathsf{XZ}|\mathsf{XY}}(x, z|x', y) = [\![x = x']\!]\mathbb{K}^{\mathsf{Z}|\mathsf{XY}}(z|x', y) \tag{154}$$

$$\tag{155}$$

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

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

$$\mathbb{K}^{\mathsf{XZ}|\mathsf{XY}}(x', z|x, y) = 0 = [\![x = x']\!]\mathbb{L}(z|x, y) \tag{156}$$

$$\tag{157}$$

But then

$$\mathbb{K}^{\mathsf{Z}|\mathsf{XY}}(y, z|x) = \sum_{x' \in X} \mathbb{K}^{\mathsf{XZ}|\mathsf{XY}}(x', z|x, y) \tag{158}$$

$$= \mathbb{L}(z|x, y) \tag{159}$$

$$\tag{160}$$

$\square$

Theorem 4.14 (Existence of valid conditional probabilities)**.** *Given a probability gap model $\mathbb{P}_\square : A \to \Delta(\Omega)$ along with a valid conditional probability $\mathbb{P}_\square^{\mathsf{XY}|\mathsf{W}}$, there exists a valid conditional probability $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{WX}}$.*

Proof. From Lemma 2.14, we have the existence of some Markov kernel $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{WX}} : W \times X \to Y$ such that

$$\mathbb{P}_\square^{\mathsf{XY}|\mathsf{W}} = \mathbb{P}_\square^{\mathsf{X}|\mathsf{W}} \odot \mathbb{P}_\square^{\mathsf{Y}|\mathsf{WX}} \tag{161}$$

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

$$\mathbb{P}_\alpha^{\mathsf{WXY}} = \mathbb{P}_\alpha^{\mathsf{W}} \odot \mathbb{P}_\square^{\mathsf{XY}|\mathsf{W}} \tag{162}$$

Thus

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{WXY}} &= \mathbb{P}_\alpha^{\mathsf{W}} \odot (\mathbb{P}_\square^{\mathsf{X}|\mathsf{W}} \odot \mathbb{P}_\square^{\mathsf{Y}|\mathsf{WX}}) &(163)\\ &= (\mathbb{P}_\alpha^{\mathsf{W}} \odot \mathbb{P}_\square^{\mathsf{X}|\mathsf{W}}) \odot \mathbb{P}_\square^{\mathsf{Y}|\mathsf{WX}}) &(164)\end{aligned}$$

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

$$(\mathsf{W}, \mathsf{X}) = (\mathrm{idf}_{W\times X} \otimes \mathrm{erasef}_Y) \circ (\mathsf{W}, \mathsf{X}, \mathsf{Y}) \tag{165}$$

By Lemma **??** together with Theorem 2.18 we have for all α:

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{XW}} &= \mathbb{P}_\alpha^{\mathsf{WXY}}(\mathrm{id}_{W\times X} \otimes \mathrm{erase}_Y) &(166)\\ &= \mathbb{P}_\alpha^{\mathsf{W}} \odot (\mathbb{P}_\square^{\mathsf{X}|\mathsf{W}} \odot \mathbb{P}_\square^{\mathsf{Y}|\mathsf{WX}})(\mathrm{id}_{W\times X} \otimes \mathrm{erase}_Y) &(167)\\ &= \mathbb{P}_\alpha^{\mathsf{W}} \odot \mathbb{P}_\square^{\mathsf{X}|\mathsf{W}} &(168)\end{aligned}$$

Then

$$\mathbb{P}_\alpha^{\mathsf{XWY}} = (\mathbb{P}_\alpha^{\mathsf{XW}}) \odot \mathbb{P}_\square^{\mathsf{Y}|\mathsf{WX}}) \tag{169}$$

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

4.5 Validity

Theorem 4.15 (Validity)**.** *Given* $(\Omega, \mathcal{F})$*,* $\mathsf{X} : \Omega \to X$*,* $\mathbb{J} \in \Delta(X)$ *with* Ω *and* X *standard measurable, there exists some* $\mu \in \Delta(\Omega)$ *such that* $\mu^{\mathsf{X}} = \mathbb{J}$ *if and only if* $\mathbb{J}$ *is a valid 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 $\mathsf{X} \bowtie x = \emptyset$ but $\mathbb{J}(x) > 0$. Then

$$\mu^{\mathsf{X}}(x) = \mu(\mathsf{X} \bowtie x) \tag{170}$$

$$= \sum_{x' \in X} \mathbb{J}(x')\mathbb{K}(\mathsf{X} \bowtie x|x') \tag{171}$$

$$= 0 \tag{172}$$

$$\neq \mathbb{J}(x) \tag{173}$$

□

Theorem 4.16 (Copy-product is an intersection of probability sets)**.** *Given* $(\Omega, \mathcal{F})$*,* $\mathsf{X} : \Omega \to (X, \mathcal{X})$*,* $\mathsf{Y} : \Omega \to (Y, \mathcal{Y})$*,* $\mathsf{Z} : \Omega \to (Z, \mathcal{Z})$ *all standard measurable and valid candidate conditionals* $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ *and* $\mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}$ *defining probability sets* $\mathbb{P}_{\{\}}$ *and* $\mathbb{Q}_{\{\}}$*, then the probability set* $\mathbb{R}_{\{\}}$ *defined by* $\mathbb{R}_{\{\}}^{\mathsf{YZ}|\mathsf{X}} := \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}$ *is equal to* $\mathbb{P}_{\{\}} \cap \mathbb{Q}_{\{\}}$*.*

Proof. By assumption

$$\mathbb{R}_{\{\}}^{\mathsf{YZ}|\mathsf{X}} := \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{YX}} \tag{174}$$

Therefore for any $\mathbb{R}_a \in \mathbb{R}_{\{\}}$

$$\mathbb{R}_a^{\mathsf{XYZ}} = \mathbb{R}_a^{\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{YX}} \tag{175}$$

$$\implies \mathbb{R}_a^{\mathsf{XY}} = \mathbb{R}_a^{\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \tag{176}$$

$$\wedge \mathbb{R}_a^{\mathsf{XYZ}} = \mathbb{R}_a^{\mathsf{XY}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{YX}} \tag{177}$$

Thus $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ is a version of $\mathbb{R}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ and $\mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}$ is a version of $\mathbb{R}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}$ 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 by definition of $\mathbb{R}_{\{\}}$, $\mathbb{R}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$ is not a version of $\mathbb{S}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$. But by construction of $\mathbb{S}$, $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ is a version of $\mathbb{S}^{\mathsf{Y}|\mathsf{X}}$ and $\mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}$ is a version of $\mathbb{S}^{\mathsf{Z}|\mathsf{YX}}$. But then by the definition of disintegration, $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}$ is a version of $\mathbb{S}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$ and so $\mathbb{R}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$ is a version of $\mathbb{S}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$, a contradiction. □

Lemma 4.17 (Equivalence of validity definitions)**.** *Given* $\mathsf{X} : \Omega \to X$*, with* Ω *and* X *standard measurable, a probability measure* $\mathbb{P}^{\mathsf{X}} \in \Delta(X)$ *is valid if and only if the conditional* $\mathbb{P}^{\mathsf{X}|*} := * \mapsto \mathbb{P}^{\mathsf{X}}$ *is valid.*

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

$$\forall A \in \mathcal{X} : \mathsf{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathsf{X}|*}(A|*) = 0 \tag{178}$$

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

$$\forall A \in \mathcal{X} : \mathsf{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathsf{X}}(A) = 0 \tag{179}$$

□

Lemma 4.18 (Copy-product of valid candidate conditionals is valid)**.** *Given* $(\Omega, \mathcal{F})$*,* $\mathsf{X} : \Omega \to X$*,* $\mathsf{Y} : \Omega \to Y$*,* $\mathsf{Z} : \Omega \to Z$ *(all spaces standard measurable) and any valid candidate conditional* $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ *and* $\mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}$*,* $\mathbb{P}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}$ *is also a valid candidate conditional.*

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

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

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

$$\mathbb{R}^{\mathsf{YZ}|\mathsf{X}}(B \times C|x) = \int_B \mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}(C|y,x)\mathbb{P}^{\mathsf{Y}|\mathsf{X}}(dy|x) \tag{180}$$

$$\leq \mathbb{P}^{\mathsf{Y}|\mathsf{X}}(B|x) \tag{181}$$

$$= 0 \tag{182}$$

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

$$\mathbb{R}^{\mathsf{YZ}|\mathsf{X}}(B \times C|x) = \int_B \mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}(C|y,x)\mathbb{P}^{\mathsf{Y}|\mathsf{X}}(dy|x) \tag{183}$$

$$= 0 \tag{184}$$

□

Corollary 4.19 (Valid conditionals are validly extendable to valid distributions)**.** *Given* Ω*,* $\mathsf{U} : \Omega \to U$*,* $\mathsf{W} : \Omega \to W$ *and a valid candidate conditional* $\mathbb{T}^{\mathsf{W}|\mathsf{U}}$*, then for any valid candidate conditional* $\mathbb{V}^{\mathsf{U}}$*,* $\mathbb{V}^{\mathsf{U}} \odot \mathbb{T}^{\mathsf{W}|\mathsf{U}}$ *is a valid candidate probability.*

Proof. Applying Lemma 4.18 choosing $\mathsf{X} = *$, $\mathsf{Y} = \mathsf{U}$, $\mathsf{Z} = \mathsf{W}$ and $\mathbb{P}^{\mathsf{Y}|\mathsf{X}} = \mathbb{V}^{\mathsf{U}|*}$ and $\mathbb{Q}^{\mathsf{Z}|\mathsf{YX}} = \mathbb{T}^{\mathsf{W}|\mathsf{U}*}$ we have $\mathbb{R}^{WU|*} := \mathbb{V}^{\mathsf{U}|*} \odot \mathbb{T}^{\mathsf{W}|\mathsf{U}*}$ is a valid conditional probability. Then $\mathbb{R}^{\mathsf{WU}} \cong \mathbb{R}^{\mathsf{WU}|*}$ is valid by Theorem 4.17. □

Theorem 4.20 (Validity of conditional probabilities)**.** *Suppose we have Ω, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, with Ω, X, Y discrete. A conditional $\mathbb{T}^{\mathsf{Y}|\mathsf{X}}$ is valid if and only if for all valid candidate distributions $\mathbb{V}^{\mathsf{X}}$, $\mathbb{V}^{\mathsf{X}} \odot \mathbb{T}^{\mathsf{Y}|\mathsf{X}}$ is also a valid candidate distribution.*

Proof. If: this follows directly from Corollary 4.19.

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

$$
\begin{aligned}
(\mathbb{V}^{\mathsf{X}} \odot \mathbb{T}^{\mathsf{Y}|\mathsf{X}})(x, y) &= \mathbb{T}^{\mathsf{Y}|\mathsf{X}}(y|x)\mathbb{V}^{\mathsf{X}}(x) && (185)\\
&= \mathbb{T}^{\mathsf{Y}|\mathsf{X}}(y|x) && (186)\\
&> 0 && (187)
\end{aligned}
$$

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

4.6 Combs

Theorem 2.32 (Equivalence of comb representations)**.** *Given sample space $(\Omega, \mathcal{F})$, a finite collection of variables $\mathsf{X}_i : \Omega \to (X_i, \mathcal{X}_i)$ for $i \in [n]$, X_i discrete, and a disassembled probability comb $\{\mathbb{P}_{\square}^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in \mathsf{X}_{[n]_{odd}}\}$, for any $l \in [n]_{\mathrm{odd}}$ and any $\mathbb{K} : X_{[l-1]} \to X_l$*

$$
\begin{aligned}
(\underline{\bigodot}_{j\in[l-1]_{\mathrm{odd}}} \mathbb{P}_{\square}^{\mathsf{X}_j|\mathsf{X}_{[j-1]}})\underline{\odot}\mathbb{K} &\overset{\mathbb{P}_\square}{\cong} (\underline{\bigodot}_{j\in[l]_{\mathrm{odd}}} \mathbb{P}_{\square}^{\mathsf{X}_j|\mathsf{X}_{[j-1]}}) && (188)\\
\implies \mathbb{K} &\overset{\mathbb{P}_\square}{\cong} \mathbb{P}_{\square}^{\mathsf{X}_l|\mathsf{X}_{[l-1]}} && (189)
\end{aligned}
$$

Proof. Equality is trivial for $l = 1$.

For a sequence $x_{[l-2]} \in X_{[l-2]}$, let $e_{[l-2]}$ be the even indices of $x_{[l-2]}$ and $o_{[l-2]}$ be the odd indices.

For any $e_{l-1} \in X_{l-1}$, $A \in X_l$ let $C^{>}_{A,e_{l-1}} \in \mathcal{X}_{[l-2]}$ be the set of points $C^{>}_{A,e_{l-1}} := \{x_{[l-2]} | \mathbb{K}(A|e_{l-1}, x_{[l-2]}) > \mathbb{P}_{\square}^{\mathsf{X}_j|\mathsf{X}_{[j-1]}}(A|e_{l-1}, x_{[l-2]})\}$, and $C^{<}_{A,e_{l-1}}$ the obvious analog. Then, definining $C_{A,e_{l-1}} = C^{>}_{A,e_{l-1}} \cup C^{<}_{A,e_{l-1}}$,

$$
\begin{aligned}
&(\underline{\bigodot}_{j\in[l-1]_{\mathrm{odd}}} \mathbb{P}_{\square}^{\mathsf{X}_j|\mathsf{X}_{[j-1]}})(A \times C^{>}_{A,e_{l-1}} | e_{[l-3]}, e_{l-1}) = 0 && (190)\\
&(\underline{\bigodot}_{j\in[l-1]_{\mathrm{odd}}} \mathbb{P}_{\square}^{\mathsf{X}_j|\mathsf{X}_{[j-1]}})(A \times C^{<}_{A,e_{l-1}} | e_{[l-3]}, e_{l-1}) = 0 && (191)\\
\implies &(\underline{\bigodot}_{j\in[l-1]_{\mathrm{odd}}} \mathbb{P}_{\square}^{\mathsf{X}_j|\mathsf{X}_{[j-1]}})(A \times C_{A,e_{l-1}} | e_{[l-3]}, e_{l-1}) = 0 && (192)\\
&= \sum_{o_{[l-2]} \in C_{A,e_{l-1},\mathrm{odd}}} \mathbb{K}(A|e_{l-1}, x_{[l-2]}) \prod_{j\in[l-2]_{\mathrm{odd}}} \mathbb{P}_{\square}^{\mathsf{X}_j|\mathsf{X}_{[j-1]}}(o_j && (193)
\end{aligned}
$$

for all $e_{[l-3]} \in X_{[l-3]_{\text{even}}}$.

Consider arbitrary $\mathbb{P}_\alpha \in \mathbb{P}_\square$, $A \subset X_l$, $C \subset X_{[l-1]}$:

$$\mathbb{P}_\alpha^{\mathsf{X}_{[l]}}(A \times C) = \sum_{x_{[l-2]} \in C} \mathbb{P}_\square(A|e_{l-1}, x_{[l-2]}) \prod_{j \in [l-2]_{\text{odd}}} \mathbb{P}_\square^{\mathsf{X}_j|\mathsf{X}_{[j-1]}}(x_j|o_{[j-2]}, e_{[j-1]}) \mathbb{P}_\alpha^{\mathsf{X}_{j+1}|\mathsf{X}_{[j]}}(x_{j+1}|o_{[j]}, e_{[j-1]}) \tag{194}$$

$$= \sum_{x_{[l-2]} \in C_{A,e_{l-1}}} \mathbb{P}_\square(A|e_{l-1}, x_{[l-2]}) \prod_{j \in [l-2]_{\text{odd}}} \mathbb{P}_\square^{\mathsf{X}_j|\mathsf{X}_{[j-1]}}(x_j|o_{[j-2]}, e_{[j-1]}) \mathbb{P}_\alpha^{\mathsf{X}_{j+1}|\mathsf{X}_{[j]}}(x_{j+1}|o_{[j]}, e_{[j-1]}) \tag{195}$$

$$+ \sum_{x_{[l-2]} \in C^C_{A,e_{l-1}}} \mathbb{P}_\square(A|e_{l-1}, x_{[l-2]}) \prod_{j \in [l-2]_{\text{odd}}} \mathbb{P}_\square^{\mathsf{X}_j|\mathsf{X}_{[j-1]}}(x_j|o_{[j-2]}, e_{[j-1]}) \mathbb{P}_\alpha^{\mathsf{X}_{j+1}|\mathsf{X}_{[j]}}(x_{j+1}|o_{[j]}, e_{[j-1]}) \tag{196}$$

$$= 0 + \sum_{x_{[l-2]} \in C^C_{A,e_{l-1}}} \mathbb{K}(A|e_{l-1}, x_{[l-2]}) \prod_{j \in [l-2]_{\text{odd}}} \mathbb{P}_\square^{\mathsf{X}_j|\mathsf{X}_{[j-1]}}(x_j|o_{[j-2]}, e_{[j-1]}) \mathbb{P}_\alpha^{\mathsf{X}_{j+1}|\mathsf{X}_{[j]}}(x_{j+1}|o_{[j]}, e_{[j-1]}) \tag{197}$$

$$= \mathbb{P}_\alpha^{\mathsf{X}_{[l-1]}} \odot \mathbb{K}(A \times C) \tag{198}$$

$$\implies \mathbb{K} \overset{\mathbb{P}_\square}{\cong} \mathbb{P}_\square^{\mathsf{X}_l|\mathsf{X}_{[l-1]}} \tag{199}$$

□

4.7 Comb conditional correspondence

Theorem 2.36 (Comb-conditional correspondence)**.** *Given a probability comb* $\{\mathbb{P}_\square^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in \mathsf{X}_{D_{odd}}\}$ *and a blind choice* α

$$\mathbb{P}_\square^{\mathsf{X}_{D_{odd}} \wr \mathsf{X}_{D_{even}}} \cong \mathbb{P}_\alpha = \mathbb{P}_\alpha^{\mathsf{X}_{D_{odd}}|\mathsf{X}_{D_{even}}} \tag{200}$$

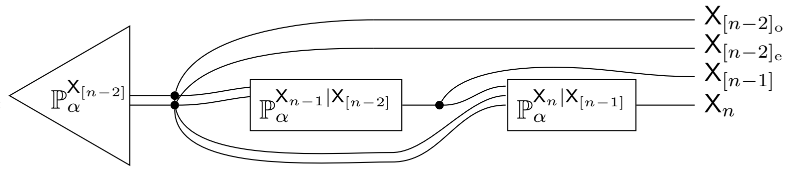
Proof. Consider $n \in D$. The correspondence is immediate for $n = 1$:

$$\mathbb{P}_\square^{\mathsf{X}_1 \wr \mathsf{X}_0} \overset{\mathbb{P}_\alpha}{\cong} \mathbb{P}_\alpha^{\mathsf{X}_1|\mathsf{X}_0} \tag{201}$$

Suppose for induction the correspondence holds for odd $n-2$. For any blind

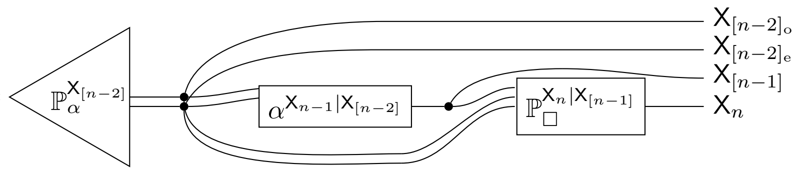
α

$$\mathbb{P}_\alpha^{\mathsf{X}_{[n]}|\mathsf{X}_0} =$$



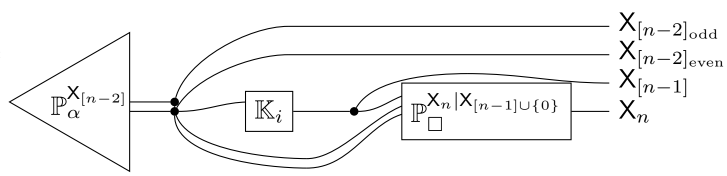
(202)

$$=$$



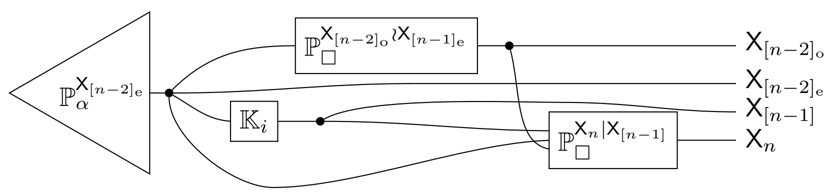
(203)

$$=$$



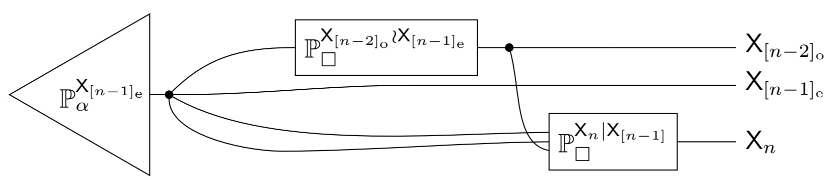
(204)

$$=$$



(205)

$$=$$



(206)

and we also have

$$(\mathbb{P}_\alpha^{\mathsf{X}_{[n]_\mathrm{even}}|\mathsf{X}_0} \odot \mathbb{P}_\square^{\mathsf{X}_{D_\mathrm{odd}} \wr \mathsf{X}_{D_\mathrm{even}}})(A \times B|x) = \tag{207}$$

$\square$

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.

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. P. Dawid. Causal Inference without Counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424, June 2000. ISSN 0162-1459. doi: 10.1080/01621459.2000.10474210. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.2000.10474210. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.2000.10474210.

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

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.

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 Mikoaj Bojaczyk and Alex Simpson, editors, *Foundations of Software Science and Computation Structures*, Lecture Notes in Computer Science, pages 313–329. Springer International Publishing, 2019. ISBN 978-3-030-17127-8.

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

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

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

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

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

Karl Menger. Random Variables from the Point of View of a General Theory of Variables. In Karl Menger, Bert Schweizer, Abe Sklar, Karl Sigmund, Peter Gruber, Edmund Hlawka, Ludwig Reich, and Leopold Schmetterer, editors, *Selecta Mathematica: Volume 2*, pages 367–381. Springer, Vienna,

2003. ISBN 978-3-7091-6045-9. doi: 10.1007/978-3-7091-6045-9_31. URL https://doi.org/10.1007/978-3-7091-6045-9_31.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009.

Judea Pearl. Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes. *Journal of Causal Inference*, 6(2), 2018. doi: 10.1515/jci-2018-2001. URL https://www.degruyter.com/view/j/jci.2018.6.issue-2/jci-2018-2001/jci-2018-2001.xml.

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.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

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.

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: