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

David Johnston

February 10, 2022

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).

Causal inference work undertaken using either approaches is often directed towards determining the likely effects of different actions that could be taken. This kind of application is strongly suggested by the terminology of "interventions" and "potential outcomes". However, if we want to reason clearly about using data to inform choices of actions, suggestive terminology is not enough to underpin a sound understanding of the correspondence between causal inference models and action selection problems.

As a motivating example, Hernán and Taubman (2008) observed that many epidemiological papers have been published estimating the "causal effect" of body mass index. However, Hernán argued, because there are many different *actions* that might affect body mass index, the potential outcomes associated with body mass index themselves are ill-defined. This would not be particularly problematic if we regarded the search for treatment effects as an endeavour entirely separate from questions of choosing actions – it's only because we want potential outcomes to tell us something about effects of actions that a many-to-one relationship between "actions" and "causal variables" becomes troublesome.

In a response to Hernán's observation, Pearl (2018) argues that *interventions* (by which we mean the operation defined by a causal graphical model) are well defined, but by default they describe "virtual interventions" or "ideal, atomic interventions", and more complicated intervention operations may be needed for a good models of real actions. In this view, the relationship between interventions and actions is clearly not straightforward. In particular, one might wonder what standard we can use to determine if an action is "ideal" and "atomic" (apart from the question begging standard of agreement with interventions in a given causal graphical model).

In another response, Shahar (2009) argued that a properly specified intervention on body mass index will necessarily yield a conclusion that intervention on body mass index has no effect at all on anything apart from body mass index itself. If this is accepted, then it might seem that there is a whole body of literature devoted to estimating a "causal effect" that is necessarily equal to zero! It seems that there is a need to clarify the relationship between actions and causal effects.

1.1 Our approach

We focus our attention on the following problem: given an experiment with sequence of variable pairs $(\mathsf{X}_i, \mathsf{Y}_i)$ and a collection of decision functions A that the experimenter may choose, when is there a unique probabilistic function $H \times X \nrightarrow Y$ that defines the "causal consequence" of X_i on Y_i for all i? We choose the setting of repeatable experiments because causal inferences are, in practice, usually drawn from sequential data generated from repeatable experiments or sampling procedures. Here the set H represents a set of hypotheses that – under some choices of decision function – becomes deterministic in the limit of infinite data.

To answer this question, we require clarity on what we mean by "variable", and so we begin with an explanation of a theory of variables. This theory is close to a standard account, but we are somewhat more explicit than usual about the relationship between variables and measurement procedures.

We then address the problem of creating probabilistic models of variables that permit us to evaluate different choices of decision function. To this end, we introduce *probability sets*, which can be though of as partially specified probability models, and *probability gap models* which can be thought of as probability sets along with a selection of choices for "filling the gap". We typically specify probability sets with conditional probabilities, and the criterion of *validity* ensures that we don't inadvertently end up specifying empty probability sets – a problem that can arise in the context of modelling interventions on body mass index.

We then prove that a condition we call *causal contractibility* is equivalent to the existence of repeatable causal consequences. This result is akin to De Finetti's theorem showing the equivalence between exchangeability and the existence of a "unique but unknown" distribution such that all variables are independent and identically distributed according to it.

Finally, we consider the question of *when* casual contractibility might be reasonable to assume. This requires us to consider the measurement processes associated with the variables that we think may or may not be causally contractible. We suggest two sufficient sets of conditions:

1. The X_is are deterministically equal for all experimental units
2. The evidence is symmetric for all experimental units, the order of experimentation is irrelevant and the X_i are deterministic given the choice of decision function

3. There exists some $(\mathsf{D}_i, \mathsf{Y}_i)_{i \in M}$ causally contractible and $\mathsf{Y}_i \perp\!\!\!\perp \mathsf{D}_i | \mathsf{HX}_i$

The first two conditions for causal contractibility are very unlikely to hold for body mass index, both because body mass index may not be functionally related to the available actions and because body mass index cannot be deterministically controlled. The third condition might hold for body mass index with respect to some sets of actions, but this is an empirical question.

2 Probability

2.1 Section outline

following section hasn't been revised

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 **??**, 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.

end previous section hasn't been revised

2.2 Standard probability theory

Definition 2.1 (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 2.2 (Markov kernel)**.** Given measure spaces $(X, \mathcal{X})$, $(Y, \mathcal{Y})$ $\mathsf{Y} : \Omega \to Y$, a Markov kernel $\mathbb{Q} : X \nrightarrow Y$ 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 2.3 (Delta measure)**.** Given a measureable 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 2.4 (Probability space)**.** A probability space is a triple $(\mu, \Omega, \mathcal{F})$, where μ is a base measure on $\mathcal{F}$.

Definition 2.5 (Variable)**.** Given a measureable space $(\Omega, \mathcal{F})$ and a set of values $(X, \mathcal{X})$, an *X-valued variable* is a measurable function $\mathsf{X} : \Omega \to X$.

Definition 2.6 (Sequence of variables)**.** Given a measureable space $(\Omega, \mathcal{F})$ and two variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, $(\mathsf{X}, \mathsf{Y}) : \Omega \to X \times Y$ is the variable $\omega \mapsto (\mathsf{X}(\omega), \mathsf{Y}(\omega))$.

Definition 2.7 (Marginal distribution with respect to a probability space)**.** Given a probability space $(\mu, \Omega, \mathcal{F})$ and a 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 $A \in \mathcal{X}$.

Lemma 2.8 (Marginal distribution as a kernel product)**.** *Given a probability space $(\mu, \Omega, \mathcal{F})$ and a variable $\mathsf{X} : \Omega \to (X, \mathcal{X})$, define $\mathbb{F}_{\mathsf{X}} : \Omega \nrightarrow X$ by $\mathbb{F}_{\mathsf{X}}(A|\omega) = \delta_{\mathsf{X}(\omega)}(A)$, then*

$$\mu^{\mathsf{X}} = \mu \mathbb{F}_{\mathsf{X}} \tag{1}$$

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

$$\begin{aligned} \mu\mathbb{F}_{\mathsf{X}}(A) &= \int_{\Omega} \delta_{\mathsf{X}(\omega)}(A) \mathrm{d}\mu(\omega) && (2)\\ &= \int_{\mathsf{X}^{-1}(\omega)} \mathrm{d}\mu(\omega) && (3)\\ &= \mu^{\mathsf{X}}(A) && (4) \end{aligned}$$

□

2.3 Not quite standard probability theory

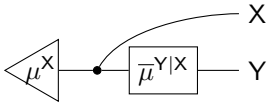
Instead of having probability distributions and Markov kernels as two different kinds of thing, we can identify probability distributions with Markov kernels whose domain is a one element set $\{*\}$.

Definition 2.9 (Probability measures as Markov kernels)**.** Given $(X, \mathcal{X})$ and $\mu \in \Delta(X)$, the Markov kernel $\mathbb{K} : \{*\} \to X$ 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 2.10 (Regular conditional distribution)**.** Given a probability space (μ, Ω) and variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, the probability of Y given X is any Markov kernel $\mu^{\mathsf{Y}|\mathsf{X}} : X \to Y$ such that

$$\mu^{\mathsf{XY}}(A \times B) = \int_A \mu^{\mathsf{Y}|\mathsf{X}}(B|x) \mathrm{d}\mu^{\mathsf{X}}(x) \qquad \forall A \in \mathcal{X}, B \in \mathcal{Y} \tag{5}$$

$$\iff \tag{6}$$

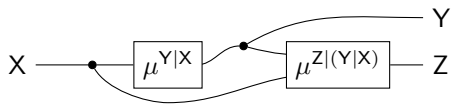
$\mu^{\mathsf{XY}} =$ 

(7)

We define higher order conditionals as "conditionals of conditionals"

Definition 2.11 (Regular higher order conditionals)**.** Given a probability space (μ, Ω) and variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and $\mathsf{Z} : \Omega \to Z$, a higher order conditional $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})} : X \times Y \to Z$ is any Markov kernel such that, for some $\mu^{\mathsf{Y}|\mathsf{X}}$,

$$\mu^{\mathsf{ZY}|\mathsf{X}}(B \times C|x) = \int_B \mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}(C|x, y) \mu^{\mathsf{Y}|\mathsf{X}}(dy|x) \tag{8}$$



$$\iff \mu^{\mathsf{ZY}|\mathsf{X}} = \tag{9}$$

Higher order conditionals are useful because $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ is a version of $\mu^{\mathsf{X}|\mathsf{YX}}$, so if we're given $\mu^{\mathsf{ZY}|\mathsf{X}}$ and we can find some $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ then we have a version of $\mu^{\mathsf{X}|\mathsf{YX}}$. This also hold for conditional with respect to probability sets, which we will introduce later (Theorem 5.4).

Furthermore, given regular $\mu^{\mathsf{XY}|\mathsf{Z}}$ and X, Y standard measurable, it has recently been proven that a regular higher order conditional $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ exists Bogachev and Malofeev (2020), Theorem 3.5. See also Theorem 5.3 for the extension of this theorem to probability sets.

2.4 Probabilistic models for causal inference

following section hasn't been revised

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.

Previous section hasn't been revised

2.5 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 2.12 (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 2.7 and 2.10 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 2.18) with respect to probability sets.

Definition 2.13 (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 $\mathbb{P}^{\mathsf{X}}_{\{\}} = \mathbb{P}^{\mathsf{X}}_{\alpha}$ for any $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$ if a distribution satisfying this condition exists. Otherwise, it is undefined.

Definition 2.14 (Regular conditional distribution 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 conditional $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\{\}}$ is any Markov kernel $X \twoheadrightarrow Y$ such that $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\{\}}$ is an $\mathsf{Y}|\mathsf{X}$ conditional probability of $\mathbb{P}_\alpha$ for all $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$. If no such Markov kernel exists, $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\{\}}$ is undefined.

Definition 2.15 (Regular higher order conditional with respect to a probability set)**.** Given a fundamental probability set Ω, variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_{\{\}}$, if $\mathbb{P}^{\mathsf{ZY}|\mathsf{X}}_{\{\}}$ exists then a higher order conditional $\mathbb{P}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}_{\{\}}$ is any Markov kenrel $X \times Y \twoheadrightarrow Z$ that is a higher order conditional of some version of $\mathbb{P}^{\mathsf{ZY}|\mathsf{X}}_{\{\}}$. If no $\mathbb{P}^{\mathsf{ZY}|\mathsf{X}}_{\{\}}$ exists, $\mathbb{P}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}_{\{\}}$ is undefined.

Under the assumption of standard measurable spaces, the existence of a conditional probability $\mathbb{P}^{\mathsf{ZY}|\mathsf{X}}_{\{\}}$ implies the existence of a higher order conditional

$\mathbb{P}_{\{\}}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ with respect to the same probability set (Theorem 5.3). $\mathbb{P}_{\{\}}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ is in turn a version of the conditional $\mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}$ (Theorem 5.4). Thus, from the existence of $\mathbb{P}_{\{\}}^{\mathsf{ZY}|\mathsf{X}}$ we can derive the existence of $\mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}$.

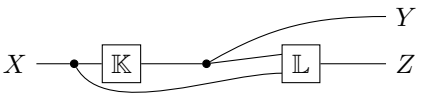
2.6 Semidirect product and almost sure equality

The operation used in Equation 7 that combines μ^{X} and $\mu^{\mathsf{Y}|\mathsf{X}}$ is something we will use repeatedly, so we call it the *semidirect product* and give it the symbol $\odot$. We also define a notion of almost sure equality with respect to $\odot$: $\mathbb{K} \overset{\mu^{\mathsf{X}}}{\cong} \mathbb{L}$ if $\mu^{\mathsf{X}} \odot \mathbb{K} = \mu^{\mathsf{X}} \odot \mathbb{L}$. Thus if two terms are almost surely equal, they are substitutable when they both appear in a semidirect product.

Definition 2.16 (Semidirect product)**.** Given $\mathbb{K} : X \nrightarrow Y$ and $\mathbb{L} : Y \times X \nrightarrow Z$, define the copy-product $\mathbb{K} \odot \mathbb{L} : X \to Y \times Z$ as

$$\mathbb{K} \odot \mathbb{L} := \mathrm{copy}_X(\mathbb{K} \otimes \mathrm{id}_X)(\mathrm{copy}_Y \otimes \mathrm{id}_X)(\mathrm{id}_Y \otimes \mathbb{L}) \tag{10}$$

$$=$$



(11)

$$\iff \tag{12}$$

$$(\mathbb{K} \odot \mathbb{L})(A \times B|x) = \int_A \mathbb{L}(B|y,x)\mathbb{K}(dy|x) \qquad A \in \mathcal{Y}, B \in \mathcal{Z} \tag{13}$$

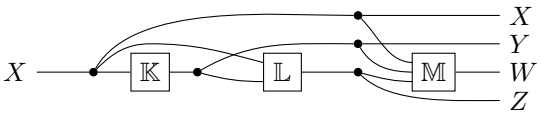
Lemma 2.17 (Semidirect product is associative)**.** *Given* $\mathbb{K} : X \nrightarrow Y$, $\mathbb{L} : Y \times X \nrightarrow Z$ *and* $\mathbb{M} : Z \times Y \times X \nrightarrow W$

$$(\mathbb{K} \odot \mathbb{L}) \odot \mathbb{Z} = \mathbb{K} \odot (\mathbb{L} \odot \mathbb{Z}) \tag{14}$$

$$\tag{15}$$

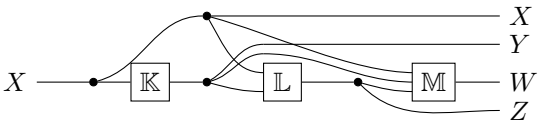
Proof.

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



(16)

$$=$$



(17)

$$= \mathbb{K} \odot (\mathbb{L} \odot \mathbb{M}) \tag{18}$$

□

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

Definition 2.18 (Almost sure equality)**.** Two Markov kernels $\mathbb{K} : X \nrightarrow Y$ and $\mathbb{L} : X \nrightarrow Y$ 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 $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$,

$$\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{K} = \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{L} \tag{19}$$

Lemma 2.19 (Conditional probabilities are almost surely equal)**.** *If $\mathbb{K} : X \nrightarrow Y$ and $\mathbb{L} : X \nrightarrow Y$ are both versions of $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ then $\mathbb{K} \overset{\mathbb{P}_{\{\}}}{\cong} \mathbb{L}$*

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

$$\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{K} = \mathbb{P}_\alpha^{\mathsf{XY}} \tag{20}$$

$$= \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{L} \tag{21}$$

□

Lemma 2.20 (Substitution of almost surely equal Markov kernels)**.** *Given $\mathbb{P}_{\{\}}$, if $\mathbb{K} : X \times Y \nrightarrow Z$ and $\mathbb{L} : X \times Y \nrightarrow Z$ are almost surely equal $\mathbb{K} \overset{\mathbb{P}_{\{\}}}{\cong} \mathbb{L}$, then for any $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$*

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{K} \overset{a.s.}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{L} \tag{22}$$

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

$$\mathbb{P}_\alpha^{\mathsf{XY}} \odot \mathbb{K} = (\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}) \odot \mathbb{K} \tag{23}$$

$$= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{K}) \tag{24}$$

$$= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{L}) \tag{25}$$

□

Lemma 2.21 (Semidirect 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*

$$\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}} = \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{XY}} \tag{26}$$

$$\tag{27}$$

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

$$\mathbb{P}_\alpha^{\mathsf{XYZ}} = \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_\alpha^{\mathsf{YZ}|\mathsf{X}} \tag{28}$$

$$= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_\alpha^{\mathsf{Z}|\mathsf{YX}}) \tag{29}$$

$$= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}) \tag{30}$$

□

2.7 Maximal probability sets and valid conditionals

So far we have defined probability sets and conditional probabilities as Markov kernels that can sometimes be derived from a probability set. We are often interested in working in the opposite direction: at the outset, we have a conditional probability and and we want to reason about the largest probability set admitting this conditional probability. We call this a *maximal probability set*.

We need to be a little bit careful when we proceed in this fashion: we can't take an arbitrary Markov kernel $\kappa : X \nrightarrow Y$ and declare it to be a conditional probability $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ for some $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$ and a maximal probability set $\mathbb{P}_{\{\}}$. The reason for this is that some collections of variables cannot have arbitrary conditional probabilities, and so $\mathbb{P}_{\{\}}$ may in fact be the empty set. We address this with the notion of validity; a *valid distribution* is a distribution associated with a particular variable that defines a nonempty set of base measures on Ω (Theorem 5.9), and *valid conditionals* are a set of conditional probabilities closed under $\odot$ and reducing to valid distributions when conditioning on a trivial variable (Lemma 5.12).

Consider, for example, $\Omega = \{0, 1\}$ with $\mathsf{X} = (\mathsf{Z}, \mathsf{Z})$ for $\mathsf{Z} := \mathrm{id}_\Omega$ and any measure $\kappa \in \Delta(\{0, 1\}^2)$ such that $\kappa(\{1\} \times \{0\}) > 0$. Note that $\mathsf{X}^{-1}(\{1\} \times \{0\}) = \mathsf{Z}^{-1}(\{1\}) \cap \mathsf{Z}^{-1}(\{0\}) = \emptyset$. Thus for any probability measure $\mu \in \Delta(\{0, 1\})$, $\mu^{\mathsf{X}}(\{1\} \times \{0\}) = \mu(\emptyset) = 0$ and so κ cannot be the marginal distribution of X for any base measure at all.

Definition 2.22 (Valid distribution)**.** Given $(\Omega, \mathcal{F})$ and a variable $\mathsf{X} : \Omega \to X$, an X-valid probability distribution is any probability measure $\mathbb{K} \in \Delta(X)$ such that $\mathsf{X}^{-1}(A) = \emptyset \implies \mathbb{K}(A) = 0$ for all $A \in \mathcal{X}$.

Definition 2.23 (Valid conditional)**.** Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ a $\mathsf{Y}|\mathsf{X}$-*valid conditional probability* is a Markov kernel $\mathbb{L} : X \nrightarrow Y$ such that:

$$\forall B \in \mathcal{Y}, x \in X : (\mathsf{X}, \mathsf{Y}) \bowtie \{x\} \times B = \emptyset \implies (\mathbb{L}(B|x) = 0) \vee (\mathsf{X} \bowtie \{x\} = \emptyset) \tag{31}$$

Definition 2.24 (Maximal probability set)**.** Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and a $\mathsf{Y}|\mathsf{X}$-valid conditional probability $\mathbb{L} : X \nrightarrow Y$ the maximal probability set $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}[M]}$ associated with $\mathbb{L}$ is the probability set such that for all $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$, $\mathbb{L}$ is a version of $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}$.

We use the notation $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}[M]}$ as shorthand to refer to the probability set $\mathbb{P}_{\{\}}$ maximal with respect to $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$.

Lemma 5.12 shows that the semidirect product of any pair of valid conditoinal probabilities is itself a valid conditional. Suppose we have some collection of $\mathsf{X}_i|\mathsf{X}_{[i-1]}$-valid conditionals $\{\mathbb{P}_i^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in [n]\}$; then recursively taking the semidirect product $\mathbb{M} := \mathbb{P}_1^{\mathsf{X}_1} \odot (\mathbb{P}_2^{\mathsf{X}_2|\mathsf{X}_1} \odot ...)$ yields a $\mathsf{X}_{[n]}$ valid distribution. Furthermore, the maximal probability set associated with $\mathbb{M}$ is nonempty.

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).

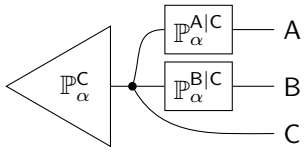
Note that validity is not a necessary condition for a conditional to define a non-empty probability set. The intuition for this is: if we have some $\mathbb{K} : X \nrightarrow Y$, $\mathbb{K}$ might be an invalid $\mathsf{Y}|\mathsf{X}$ conditional on all of X, but might be valid on some subset of X, and so we might have some probability model $\mathbb{P}$ that assigns measure 0 to the bad parts of X such that $\mathbb{K}$ is a version of $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$. On the other hand, if we want to take the product of $\mathbb{K}$ with arbitary valid X probabilities, then the validity of $\mathbb{K}$ is necessary (Theorem 5.14).

2.7.1 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 5.4.

Definition 2.25 (Conditional independence with respect to a probability model)**.** For a *probability model* $\mathbb{P}_\alpha$ 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

$$\mathbb{P}_\alpha^{\mathsf{ABC}} = \tag{32}$$



Cho and Jacobs (2019) have shown that this definition coincides with the standard notion of conditional independence for a particular probability model (Theorem 5.15).

Conditional independence satisfies the *semi-graphoid axioms*. For all standard measurable spaces $(\Omega, \mathcal{F})$ and all probability measures $\mathbb{P} \in \Delta(\Omega)$:

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 $\mathsf{A} \perp\!\!\!\perp_{\mathbb{P}} \mathsf{B}|\mathsf{W}$ 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 $\mathsf{A} \perp\!\!\!\perp_{\mathbb{P}_\square} (\mathsf{B}, \mathsf{C})|\mathsf{W}$

We define *universal conditional independence* with respect to a probability set as conditional independence for every probability model in the set.

Definition 2.26 (Universal conditional independence)**.** For a *probability set* $\mathbb{P}_{\{\}}$ and variables $\mathsf{A}, \mathsf{B}, \mathsf{Z}$, we say B is universally conditionally independent of A given C, written $\mathsf{B} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{A}|\mathsf{C}$, if for all $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$ $\mathsf{A} \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{C}|\mathsf{B}$.

It is straightforward to show that universal conditional independence satisfies the semi-graphoid axioms.

Lemma 2.27. $[\forall x : (f(x) \implies g(x))] \implies [(\forall x : f(x)) \implies (\forall x : g(x))]$

Proof.

$\forall x : f(x) \implies g(x)$		premise (33)
$\forall x : f(x)$		premise (34)
$f(a)$	universal instantiation on 34 substitute a/x	(35)
$f(a) \implies g(a)$	universal instantiation on 33 substitute a/x	(36)
$g(a)$	modus ponens 35 and 36	(37)
$\forall x : g(x)$	universal generalisation on 37	(38)
$(\forall x : f(x)) \implies (\forall x : g(x))$	conditional proof $34 - 38$	(39)
$[\forall x : (f(x) \implies g(x))] \implies [(\forall x : f(x)) \implies (\forall x : g(x))]$	conditional proof 33–39	(40)

With thanks to (https://math.stackexchange.com/users/252356/kylew) for the proof. □

Lemma 2.28. *Given a standard measurable space $(\Omega, \mathcal{F})$ and $\mathbb{P}_{\{\}}$ on Ω, universal conditional independence with respect to $\mathbb{P}_{\{\}}$ satisfies the semi-graphoid axioms.*

Proof. For a particular probability $\mathbb{P}_\alpha$, each of the semi-graphoid axioms consists of a statement of the form $\forall \mathbb{P} : f(\mathbb{P}) \implies g(\mathbb{P})$ (in the case of the first axiom, it corresponds to two such statements).

As the axioms hold for conditional independence for any probability model, we have, for arbitrary $\mathbb{P}_{\{\}}$, $\forall \mathbb{P}_\alpha \in \mathbb{P}_{\{\}} : f(\mathbb{P}_\alpha) \implies g(\mathbb{P}_\alpha)$.

Then, by Lemma 2.27, $(\forall \mathbb{P}_\alpha \in \mathbb{P}_{\{\}} : f(\mathbb{P}_\alpha)) \implies (\forall \mathbb{P}_\alpha \in \mathbb{P}_{\{\}} : g(\mathbb{P}_\alpha))$.

Note that $(\forall \mathbb{P}_\alpha \in \mathbb{P}_{\{\}} : f(\mathbb{P}_\alpha))$ is, by definition, a universal conditional independence statement with respect to $\mathbb{P}_{\{\}}$. □

3 Decision problems

We want to construct models to help make decisions. For our purposes, "making a decision" means we have some mathematically well-defined set C of choices under consideration, and some means of comparing one choice to another that

induces a partial order on choices. If we are trying to be precise with language, we might call this *formal* decision making; people may often make decisions where it is difficult to identify a set C of choices under consideration or a rule for inducing a partial order on it.

A procedure for making formal decisions is, in a sense, the opposite of a measurement procedure. With measurement, we have some unclear process that interacts with the world and leaves us with a collection of mathematical objects. Formal decision making starts with a collection of mathematical objects – a set of choices, a partial order and a tie-breaking rule which together imply a *choice*, and then on the basis of this choice some unclear procedure takes place that has consequences in the world. If a measurement is a "function" whose domain is the world, a formal decision can be thought of as a "function" whose codomain is the world.

We make the following assumptions about how choices are compared:

- Each choice is associated with a probability set over some sample space $(\Omega, \mathcal{F})$; that is, we have a function $f : C \to \mathcal{P}(\Delta(\Omega))$
- There is some method available for comparing the desirability of $\mathbb{P}_\alpha$ and $\mathbb{P}_{\alpha'}$ for all $\alpha, \alpha' \in C$. For example, we might have a utility function $u : \Omega \to \mathbb{R}$, and $|\mathbb{P}_\alpha| = 1$ for all $\alpha \in C$; then we can compare $\mathbb{P}_\alpha$ with $\mathbb{P}_{\alpha'}$ using expected utility

We can represent such a model f with probability sets. There are many different ways to do this, but the general scheme is as follows:

- There is a probability set $\mathbb{P}_\square$ representing "properties that hold regardless of which choice is taken"
- There is a collection of probability sets $\{\mathbb{P}_{\tilde{\alpha}}\}_A$ representing "properties that hold just for the choice α"
- $\mathbb{P}_\alpha = \mathbb{P}_\square \cap \mathbb{P}_{\tilde{\alpha}}$

It is always possible to accomplish this: take $\mathbb{P}_\square \supset \cup_{\alpha \in C} \mathbb{P}_\alpha$ and $\mathbb{P}_{\tilde{\alpha}} = \mathbb{P}_\alpha$. However, we might be motivated to make different choices for $\mathbb{P}_\square$ and the $\mathbb{P}_{\tilde{\alpha}}$s.

The probability set representation allows us to make use of universal conditional probabilities and universal conditional independence to reason about decision problem models. By construction $\mathbb{P}_\alpha \subset \mathbb{P}_\square$ for all α, any universal conditional independence that holds for $\mathbb{P}_\square$ holds for all $\mathbb{P}_\alpha$ as well, and similarly any universal conditional probability with respect to $\mathbb{P}_\square$ is also a universal conditional probability for all $\mathbb{P}_\alpha$.

We call this general scheme a *probability gap model*. A probability gap model specifies some universal behaviour via $\mathbb{P}_\square$, but this specification is incomplete; it has some gaps. We then have a selection of probability sets $\{\mathbb{P}_{\tilde{\alpha}}\}_A$ that specify choice-specific behaviour; this set represents different ways to fill the gap.

Definition 3.1 (Probability gap model)**.** Given $(\Omega, \mathcal{F})$, a probability gap model is a triple $(\mathbb{P}_\square, \{\mathbb{P}_{\tilde{\alpha}}\}_A, f)$ where $\mathbb{P}_\square$ is a probability set and $\{\mathbb{P}_{\tilde{\alpha}}\}_A$ is a collection of probability sets and $f : A \to \mathcal{P}(\Delta(\Omega))$ is the map

$$\mathbb{P}_\alpha := f(\alpha) \tag{41}$$

$$= \mathbb{P}_\square \cap \mathbb{P}_{\tilde{\alpha}} \tag{42}$$

3.1 Conditional probability models

A simple but interesting class of probability gap model is the *conditional probability model*. Conditional probability models can be thought of as describing problems in which there is a variable X whose marginal probability can be chosen and a variable Y that responds in a fixed way to X.

Definition 3.2 (Conditional probability model)**.** Given $(\Omega, \mathcal{F})$, $\mathsf{Y} : \Omega \to Y$, $\mathsf{X} : \Omega \to X$, a $\mathsf{Y}|\mathsf{X}$ conditional probability model is a probability gap model $(\mathbb{P}_\square^{\mathsf{Y}|\mathsf{X}[M]}, \{\mathbb{P}_\alpha^{\mathsf{X}[M]}\}_A, f)$.

Conditional probability models arise when we have variables that represent the choice made, and each choice is associated with a *unique* probability distribution over an outcome of interest.

Example 3.3 (Choice variable)**.** Suppose we have a procedure $\mathcal{C}$ that compares the elements of a countable set C and produces a choice according to some notion of the desirability of each one. Another procedure $\mathcal{Y}$ measures outcomes of interest. These are modelled with variables $\mathsf{C} : \Omega \to C$ and $\mathsf{Y} : \Omega \to Y$ respectively. Without loss of generality, we take $\Omega = C \times Y$ and C and Y are projections onto the matching sets.

For each $\alpha \in C$, suppose that $\mathbb{P}_\alpha^{\mathsf{C}}$ exists and is equal to δ_α. This expresses the notion that, if the choice procedure actually yields α, then the model should always assign probabilty 1 to $\mathsf{C} \bowtie \alpha$.

Suppose also that $\mathbb{P}_\alpha^{\mathsf{Y}}$ exists for all α. That is, the marginal probability of Y is unique given any choice α.

We thus have a model $f : C \to \Delta(\Omega)$ given by $\alpha \mapsto \mathbb{P}_\alpha^{\mathsf{CY}}$ where

$$\mathbb{P}_\alpha^{\mathsf{CY}}(B \times D) = \delta_\alpha(B)\mathbb{P}_\alpha^{\mathsf{Y}}(D) \tag{43}$$

Validity of $\mathbb{P}_\alpha^{\mathsf{CY}}$ is guaranteed by the definitions of C and Y because (C, Y) is surjective.

We can implement f with a conditional probability model $(\mathbb{P}_\square^{\mathsf{Y}|\mathsf{C}[M]}, \{\mathbb{P}_\alpha^{\mathsf{C}[M]}\}_A, f)$ where $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{C}} := (D|\alpha) \mapsto \mathbb{P}_\alpha^{\mathsf{Y}}(D)$.

Then

$$\mathbb{P}_\alpha^{\mathsf{CY}} = \mathbb{P}_\alpha^{\mathsf{C}} \odot \mathbb{P}_\square^{\mathsf{Y}|\mathsf{C}} \tag{44}$$

and so by Lemma 5.10

$$f(\alpha) = \mathbb{P}_\alpha \cap \mathbb{P}_\square \tag{45}$$

find a home for the following remarks

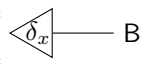
If the conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\{\}}$ and all the marginal probabilities $\mathbb{P}^{\mathsf{X}}_{\alpha}$ are valid, then by Lemma 5.12 $\mathbb{P}_{\{\}} \cap \alpha \neq \emptyset$ for all $\alpha \in A$. Thus validity of all the individual parts is enough to ensure compatibility.

We can define more complex probability gap models with a similar approach where, for example, the model is specified by an incomplete collection of conditional probabilities and the choices are each a complementary collection of conditional probabilities; we call such models *probability comb models* after Chiribella et al. (2008); Jacobs et al. (2019), but we will not address them in this paper.

3.2 Example: invalidity

Body mass index is defined as a person's weight divided by the square of their height. Suppose we have a measurment process $\mathcal{S} = (\mathcal{W}, \mathcal{H})$ and $\mathcal{B} = \frac{\mathcal{W}}{\mathcal{H}^2}$ - i.e. we figure out someone's body mass index first by measuring both their height and weight, and then passing the result through a function that divides the second by the square of the first. Thus, given the random variables W, H modelling $\mathcal{W}, \mathcal{H}$, $\mathcal{B}$ is the function given by $\mathsf{B} = \frac{\mathsf{W}}{\mathsf{H}^2}$. Given $x \in \mathbb{R}$, consider the conditional probability

$$\nu^{\mathsf{B}|\mathsf{WH}} = \tag{46}$$



Then pick some $w, h \in \mathbb{R}$ such that $\frac{w}{h^2} \neq x$ and $(\mathsf{W}, \mathsf{H}) \bowtie (w, h) \neq \emptyset$ (our measurement procedure could possibly yield (w, h) for a person's height and weight). We have $\nu^{\mathsf{B}|\mathsf{WH}}(x|w, h) = 1$, but

$$(\mathsf{B}, \mathsf{W}, \mathsf{H}) \bowtie \{(x, w, h)\} = \{\omega | (\mathsf{W}, \mathsf{H})(\omega) = (w, h), \mathsf{B}(\omega) = \frac{w}{h^2}\} \tag{47}$$

$$= \emptyset \tag{48}$$

so $\nu^{\mathsf{B}|\mathsf{WH}}$ is invalid, and there is some valid μ^{X} such that the probability set $\mathbb{P}_{\{\}}$ with $\mathbb{P}^{\mathsf{XY}}_{\{\}} = \mu^{\mathsf{X}} \odot \nu^{\mathsf{Y}|\mathsf{X}}$ is empty.

Validity rules out conditional probabilities like 46. We guess that in many cases this condition may either be trivial or unconsiously taken into account when constructing conditional probabilities. However, if we are not cognizant of the conditional our model depends on, we may inadvertently propose a model that depends on invalid conditional probabilities. For example, the conditional probability 46 would be used to evaluate the causal effect of body mass index in the causal diagram found in Shahar (2009), presuming the author used the

term "causal effect" to depend somehow on the function $x \mapsto P(\cdot|do(\mathsf{B} = x))$ as is the usual convention when discussing causal Bayesian networks.

End find a home for the following remarks

3.3 Response conditionals

Given any probability gap model $(\mathbb{P}_\square, \{\mathbb{P}_{\tilde{\alpha}}\}_A, f)$ and variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, if we have a conditional probability $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{X}}$, then this must be the conditional probability of Y given X for any $\alpha \in A$. It must therefore be the case that

$$\mathbb{P}_\alpha^{\mathsf{Y}} = \mathbb{P}_\alpha^{\mathsf{X}} \mathbb{P}_\square^{\mathsf{Y}} \qquad \forall \alpha \in A \tag{49}$$

If it is possible to control X to some extent (i.e. $\mathbb{P}_\alpha^{\mathsf{X}} \neq \mathbb{P}_{\alpha'}^{\mathsf{X}}$ for some α, α'), then $\mathbb{P}_\square^{\mathsf{Y}}$ tells us how $\mathbb{P}_\alpha^{\mathsf{X}}$ determines its effect on $\mathbb{P}_\alpha^{\mathsf{Y}}$. This feature motivates the name *response conditional* for conditional probabilities of this type.

The motivating question we introduced at the beginning of this paper was "when are potential outcomes well-defined?". This is not written using the potential outcomes framework, so we cannot directly address this question. However, we can ask "when do probability gap models feature response conditionals?".

3.4 Response conditionals and potential outcomes

There is a connection between the question of when a particular conditional probability exists and when potential outcomes are wel-defined. Given any Markov kernel there is an operation akin to function currying termed "randomness pushback" that represents the kernel as the product of a probability measure and a deterministic Markov kernel. We observe that potential outcomes models share a number of features in common with a Markov kernel represented using a randomness pushback.

Unlike function currying, there are many different randomness pushbacks that represent the same Markov kernel. The interpretation of potential outcomes models seems to require that exactly one of these possible pushbacks is a genuine potential outcomes model. Several works including Dawid (2000) and Richardson and Robins (2013) have expressed the view that some of the degrees of freedom in potential outcomes models are superfluous; both propose that the degrees of freedom that can be tested using some idealised experiment are the degrees that should be kept.

Notably, the single world intervention graphs of Richardson and Robins (2013) feature an operation that splits an "intervenable" variable X into two versions, X and X', representing "the actual X" and "the unobseved value X would have taken absent intervention" respectively (this interpretation is ours, not theirs). Thus Richardson and Robins might argue that they are interested in the existence of response conditionals of the form $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{X}\mathsf{X}'}$ rather than $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{X}}$.

We do not take a position on which degrees of freedom are good or bad in a typical potential outcomes model, nor do we explore conditionals of the form $\mathbb{P}^{\mathsf{Y}|\mathsf{X}\mathsf{X}'}$. We can, however, distinguish two different questions:

- How are response conditionals represented?
- Which response conditionals are of interest?

The first question seems to be an inconsequential stylistic matter, while the second question may determine the direction of one's analysis.

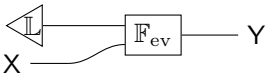
3.5 Randomness pushbacks

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 $\mathbb{K} : X \nrightarrow Y$, define $\mathbb{L} : \{*\} \nrightarrow Y^X$ by

$$\lambda\mathbb{K}((y_i)_{i\in X}) = \prod_{i\in X} \mathbb{K}(y_i|i) \tag{50}$$

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

$$\mathbb{K} = \tag{51}$$



$$\iff \tag{52}$$

$$\mathbb{K}(A|x) = \int_{Y^X} \delta_{\mathrm{ev}(y^X,x)}(A)\mathbb{L}(\mathrm{d}y^X|x) \tag{53}$$

Unlike the case of function currying, $\lambda\mathbb{K}$ is not the unique Markov kernel for which 51 holds. In fact, we can substitute any $\mathbb{M}$ such that, for any $i \in X$

$$\sum_{y_{\{i\}^C} \in Y^{|X|-1}} \mathbb{M}((y_i)_{i\in X}) = \mathbb{K}(y_i|i) \tag{54}$$

This representation of a Markov kernel is called a *randomness pushback* by Fritz (2020).

Randomness pushbacks have a few features in common with potential outcomes causal models. For our purposes, we will say a potential outcomes model is a probability set $(\Omega, \mathcal{F}, \mathbb{P}_{\{\}})$ along with variables X, Y, Y^X such that

$$\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}\mathsf{Y}^X} = \mathbb{F}_{\mathrm{ev}} \tag{55}$$

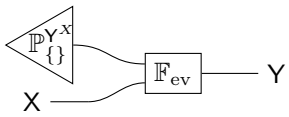
More commonly, this property is expressed as

$$\mathsf{Y} \overset{a.s.}{=} \text{ev}(\mathsf{X}, \mathsf{Y}^X) \tag{56}$$

We consider a potential outcomes model to be a probability set here, but we can formally recover a "traditional" potential outcomes model by considering probability sets of size 1.

If we additionally have the existence of $\mathbb{P}_{\{\}}^{\mathsf{Y}^X|\mathsf{X}}$ and $\mathsf{Y}^{\mathsf{X}} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{X}$ then

$$\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} = \tag{57}$$



Equation 57 is clearly a version of 51. As we have established, provided $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ exists, we can always introduce some variable Y^X and corresponding $\mathbb{P}_{\{\}}^{\mathsf{Y}^X}$ such that Equation 57 holds.

3.5.1 Choices aren't always known

One area of potentail difficulty with our approach to formalising causal inference from the starting point of modelling decision problems is related to the issue of unknown choice sets. While causal investigations are often concerned with helping someone to make better decisions, the kind of "decision making process" associated with them is not necessarily well modeled by the setup above. Often the identity of the decision maker and the exact choices at hand are vague. Consider Banerjee et al. (2016): a large scale experiment was conducted trialling a number of different strategies all aiming to increase the amount of learning level appropriate instruction available to students in four Indian states. It is not clear who, exactly, is going to make a decision on the basis of this information, but one can guess:

- They're someone with interest in and authority to make large scale changes to a school system
- They consider the evidence of effectiveness of teaching at the right level relevant to their situation
- They consider the evidence regarding which strategies work to implement this approach relevant to their situation

This could describe a writer who is considering what kind of advice they can provide in a document, a grantmaker looking to direct funds, a policy maker trying to design policies with appropriate incentives a program manager trying to implement reforms or someone in a position we haven't thought of yet. All of these people have very different choices facing them, and to some extent it is desirable that this research is relevant to all of them.

These situations are common in the field of causal inference and the question of how to formalise them may be worthy of study. The situation of known choices that we focus on here is easier to study, and it may serve as an idealisation or a limiting case of situations where choices are unknown.

3.6 Other decision theoretic causal models

Lattimore and Rohde (2019a) and Lattimore and Rohde (2019b) consider an observational probability model coupled to a collection of indexed interventional probability models, with the observational probability model coupled to the interventional models by shared (unobserved) parameters. In these papers, they show how such a model can reproduce inferences made using Causal Bayesian Networks. In our terms, Lattimore and Rohde's model family can be thought of as conditional probability models $(\mathbb{P}_{\square}^{\mathsf{OCH}|\mathsf{D}[M]}, \{\mathbb{P}_{\vec{\alpha}}^{\mathsf{D}}\}_A, f)$ where D represents the choice of intervention, O observations, I interventional consequences and $\mathbb{H}$ model parameters or hypotheses, with the properties $\mathsf{H} \perp\!\!\!\perp_{\mathbb{P}_{\square}} \mathsf{D}$, $\mathsf{O} \perp\!\!\!\perp_{\mathbb{P}_{\square}} \mathsf{D}|\mathsf{H}$ and $\mathsf{C} \perp\!\!\!\perp_{\mathbb{P}_{\square}} \mathsf{O}|\mathsf{D},\mathsf{H}$.

Note that we may well prefer to use a model in which the intervention X is chosen to depend on the observations O somehow. It's possible to represent this using probability gap models, but doing so is beyond the scope of this paper.

The approach to decision theoretic causal inference described by Dawid (2020) is somewhat different to Rohde and Lattimore's:

> 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 key difference is that Rohde and Lattimore's model employs different variables O and C to represent observations and interventional consequences, which allows us to write conditional independence statements like $\mathsf{C} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{O}|\mathsf{D},\mathsf{H}$, while Dawid's approach uses the same variable Y to represent observations and interventional consequences, depending on the choice of regime. In this scheme there is no way we can see to express something like "the observations are independent of the interventional consequences given the choice of intervention and the parameters".

If we require that the map from regimes to probability distributions is measurable (which can be trivially satisfied if the set of regimes is countable), then we can also characterise Dawid's approach using a probability set $\mathbb{P}_{\{\}}^{\mathsf{TY}|\mathsf{F}_T[M]}$ where F_T represents the prevailing regime, and T and Y are as in the quote above. Note that we do not consider conditional probability models here, because it is not obvious how we should say F_T is distributed given any choice of intervention; for any choice of intervention, F_T will sometimes take th evalue

"observational" and sometimes take a value corresponding to the chosen intervention. Moreover, it seems inappropriate to posit a sequence of independent and identically distributed copies of F_T because we are likely to know it advance exactly when it will and won't take on the observational value, and so we can't associate it with a unique marginal distribution.

Heckerman and Shachter (1995) also explore a decision theoretic approach to causal inference. Their approach is based on the decision theory of Savage (1954), and represents a decision problem in terms of choices D, outcome variables U and unobserved states of the world S; each state of the world defines a deterministic map $\mathsf{D} \twoheadrightarrow \mathsf{U}$ In comparison with the theory Lattimore and Rohde Heckerman and Schachter's comes with the added requirement of deterministic outcomes. It also features no built in distinction between observations and outcomes of interventions, though one could always consider models where U is in fact a pair of variables (O, I) such that O satisfies the independences required of observations in the Lattimore and Rohde model.

4 When do response conditionals exist?

The specific question we ask here is: given a conditional probability model $(\mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{D}[M]}, \{\mathbb{P}_{\alpha}^{\mathsf{D}}\}_A, f)$ where $\mathsf{D} = (\mathsf{D}_i)_{i \in M}$ are choice variables, $\mathsf{Y} = (\mathsf{Y}_i)_{i \in M}$ are outcome variables and M is some index set, when does there exist a response conditional $\mathbb{P}_{\square}^{\mathsf{Y}_0|\mathsf{D}_0}$ that explains how each Y_i responds to each D_i; we call these *repeatable response conditionals*. The reason why we consider a sequential problem is that in practice, causal inference almost always deals with observational data that is assumed to be appropriately modeled with a sequence of independent and identically distributed random variables. Furthermore, we can characterise precisely the kinds of models for which such response conditionals exist. Thus the sequential setting is both a theoretically tractable and widely appliccable starting point.

We examine this question from two points of view: firstly, we aks *what kind of conditional probability models exhibit repeatable response condtionals?* Secondly, we ask *what kinds of experiments are these conditional probability models appropriate for?* In the course of answering the second question, we show that apparently subtle differences in the description of an experimental procedure can determine whether a particular experiment should or should not be modeled with repeatable response conditionals.

We need to make strong assumptions about an experiment to establish the existence of repeatable response conditional in the first place, once we have repeatable response conditionals with respect to some pair of variables (Y, D), response conditionals with respect to different pairs of variables may exist due to conditional independences in the original $\mathbb{P}_{\square}^{\mathsf{Y}_0|\mathsf{D}_0}$. These conditional independences can be tested for in the observed data.

4.1 Repeatable experiments

Our setup is a conditional probability model $(\mathbb{P}_{\square}^{\overline{\mathsf{Y}|\mathsf{D}}}, A)$ where $\mathsf{Y} := \mathsf{Y}_M = (\mathsf{Y}_i)_{i\in M}$ and $\mathsf{D} := \mathsf{D}_M = (\mathsf{D}_i)_{i\in M}$ for some index set M; we say such a model is a model of a *sequential experiment*. We say that Y_i is the consequence corresponding to the decision D_i for all $i \in M$ (i.e. variables with matching indices correspond). We say that $\mathbb{P}_{\square}^{\overline{\mathsf{Y}|\mathsf{D}}}$ features *repeatable response conditionals* if there exists a hypothesis H such that $\mathbb{P}_{\square}^{\mathsf{Y}_i|\mathsf{HD}_i} = \mathbb{P}_{\square}^{\mathsf{Y}_j|\mathsf{HD}_j}$ for all $i, j \in M$, $\mathsf{H} \perp\!\!\!\perp_{\mathbb{P}_{\square}} \mathsf{D}$ and $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_{\square}} \mathsf{Y}_{M\setminus\{i\}}\mathsf{D}_{M\setminus\{i\}}|\mathsf{HD}_i$. We remind the reader that in a conditional probability model, arbitrary conditional probabilities do not always exist, see Definition 2.14.

There are two assumptions relevant to the existence of repeatable response conditionals. The first is a condition of interchangability: in particular, given a permutation of M, we get the same result from applying this permutation only to the set of actions we take, or from applying it only to the set of outcomes we observe. We call this *exchange commutativity*.

The second is a condition of *locality of consequences*; that is the assumption that Y_i is independent of D_j given D_i for any j. It is possible to have models in which commutativity to exchange holds but locality of consequences does not. Such a situation could arise in a model of stimulus payments to individuals in a nation; if exactly n payments of \$10 000 are made, we might consider that it doesn't matter much exactly who receives the payments (this is a subtle question, though, we will return to it in more detail later). However, the amount of inflation induced depends on the number of payments; making 100 such payments will have a negligible effect on inflation, while making payments to everyone in the country is likely to have a substantial effect. Dawid (2000) discusses condition of *post-treatment exchangeability* which is similarto exchange commutativity, and there he gives the example of herd immunity in vaccination campaigns as a situation where post-treatment exchangeability holds but locality of consequences does not.

As we have mentioned, exchange commutativity is similar to the condition of *post-treatment exchangeability* found in Dawid (2020). Exchange commutativity is also very similar to the exchangeability assumption of GREENLAND and ROBINS (1986), and the assumption of exchangeability found in Banerjee et al. (2017). However, in every case there is a subtle but important difference; exchange commutativity concerns exchanges of actions and outcomes, while these other exchangeability conditions concern exchange of "people" or "experimental units". Swapping people and experimental units are actions in the real world, and so these symmetries have to be described as part of the measurement procedure, and can't be purely characterised as symmetries of a probabilistic model. On the other hand, swapping the orders of variables *can* be described purely as a symmetry of a probabilistic model, as these swaps involve only function composition. As we will discuss in detail, exchangeability of experimental units does not always imply exchange commutativity.

Locality of consequences is similar to the stable unit treatment distribution

assumption (SUTDA) in Dawid (2020). It is also related to the "no interference" part of the stable unit treatment value assumption (SUTVA). The stable unit treatment value assumption (SUTVA) is given as (Rubin, 2005):

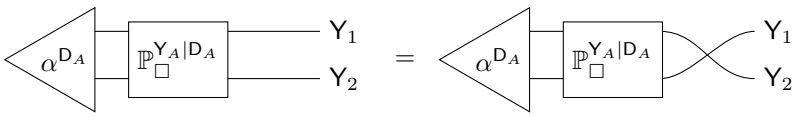
"("SUTVA) comprises two subassumptions. First, it assumes that *there is no interference between units (Cox 1958)*; that is, neither $Y_i(1)$ nor $Y_i(0)$ is affected by what action any other unit received. Second, it assumes that *there are no hidden versions of treatments*; no matter how unit i received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

Not sure if or where I want to put this, I just think it helps to illustrate the difference

Exchange commutativity is not equivalent to exchangeability in the sense of De Finetti's well-known theorem de Finetti ([1937] 1992). The latter can be understood as expressing an indifference between conducting the experiment as normal, or conducting the experiment and then swapping some labels. However, swapping *choices* will (usually) lead to different "pieces of the experiment" receiving different treatment, which is something that can't be achieved by swapping labels after the experiment has concluded.

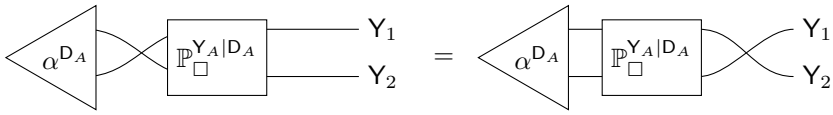
The difference is illustrated by the following pair of diagrams.

Exchangeability (swapping labels):



(58)

Exchange commutativity (swapping choices $\sim$ swapping labels):



(59)

—-end not sure where to put——

4.2 Consequence contractibility

We offer formal definitions of exchange commutativity and locality of consequences, as well as "consequence contractibility", which is the conjunction of both conditions.

A conditional probability model commutes with exchange if applying a permutation to the choice D_M "before" it is taken yields the same result as applying the corresponding permutation to Y_M "after" it is observed.

Definition 4.1 (Swap map)**.** Given $M \subset \mathbb{N}$ a finite permutation $\rho : M \to M$ and a variable $\mathsf{X} : \Omega \to X^M$ such that $\mathsf{X} = (\mathsf{X}_i)_{i \in M}$, define the Markov kernel $\mathrm{swap}_{\rho(\mathsf{X})} : X^M \nrightarrow X^M$ by $(d_i)_{i \in \mathbb{N}} \mapsto \delta_{(d_{\rho(i)})_{i \in \mathbb{N}}}$.

Definition 4.2 (Exchange commutativity)**.** Suppose we have a sample space $(\Omega, \mathcal{F})$ and a conditional probability model $(\mathbb{P}_{\Box}^{\mathsf{Y}|\mathsf{D}[M]}, \{\mathbb{P}_{\tilde{\alpha}}^{\mathsf{D}}\}_A, f)$ with $\mathsf{Y} := \mathsf{Y}_M := (\mathsf{Y}_i)_M$, $\mathsf{D} := \mathsf{D}_M := (\mathsf{D}_i)_M$, $M \subseteq \mathbb{N}$. If, for any decision rule $\alpha \in A$,

$$\mathbb{P}_{\tilde{\alpha}}^{\mathsf{D}} \odot \mathrm{swap}_{\rho(\mathsf{D})} \mathbb{P}_{\Box}^{\mathsf{Y}|\mathsf{D}} = \mathbb{P}_{\tilde{\alpha}}^{\mathsf{D}} \odot \mathbb{P}_{\Box}^{\mathsf{Y}|\mathsf{D}} \mathrm{swap}_{\rho(\mathsf{Y})} \tag{60}$$

Then $\mathbb{P}_{\Box}$ *commutes with exchange.*

A conditional probability model exhibits locality of consequences if, given two different choices that agree on an subsequence of indices, the model yields identical outcomes if we restrict our attention to the subsequence on which the different choices match. 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 $\mathbb{P}_{\tilde{\alpha}}^{\mathsf{D}_1\mathsf{D}_3} = \mathbb{P}_{\tilde{\beta}}^{\mathsf{D}_1\mathsf{D}_3}$ then $\mathbb{P}_{\alpha}^{\mathsf{Y}_1\mathsf{D}_1\mathsf{Y}_3\mathsf{D}_3} = \mathbb{P}_{\beta}^{\mathsf{Y}_1\mathsf{D}_1\mathsf{Y}_3\mathsf{D}_3}$.

Definition 4.3 (locality of consequences)**.** Suppose we have a sample space $(\Omega, \mathcal{F})$ and a conditional probability model $(\mathbb{P}_{\Box}^{\mathsf{Y}|\mathsf{D}[M]}, \{\mathbb{P}_{\tilde{\alpha}}^{\mathsf{D}}\}_A, f)$ with $\mathsf{Y} := \mathsf{Y}_M := (\mathsf{Y}_i)_M$, $\mathsf{D} := \mathsf{D}_M := (\mathsf{D}_i)_M$, $M \subseteq \mathbb{N}$. For any ordered sequence $S = (s_i)_{i \in Q}$ where $Q \subset M$ and $i < j \implies s_i < s_j$, let $\mathsf{D}_S := (\mathsf{D}_i)_{i \in S}$ and $\mathsf{D}_T := (\mathsf{D}_i)_{i \in T}$. If for any $\alpha, \beta \in R$

$$\mathbb{P}_{\alpha}^{\mathsf{D}_S} = \mathbb{P}_{\beta}^{\mathsf{D}_S} \tag{61}$$

$$\implies \mathbb{P}_{\alpha}^{(\mathsf{D}_i, \mathsf{Y}_i)_{i \in S}} = \mathbb{P}_{\beta}^{(\mathsf{D}_i, \mathsf{Y}_i)_{i \in S}} \tag{62}$$

then $\mathbb{P}_{\Box}$ exhibits *locality of consequences.*

Neither condition implies the other.

Lemma 4.4. *Exchange commutativity does not imply locality of consequences or vise versa.*

Proof. A conditional probability model that exhibits exchange commutativity but some choices have non-local consequences:

Suppose $D = Y = \{0, 1\}$ and we have a conditional probability model $(\mathbb{P}_{\Box}^{\mathsf{Y}|\mathsf{D}[M]}, \{\mathbb{P}_{\tilde{\alpha}}^{\mathsf{D}}\}_A, f)$ where $\mathsf{D} = (\mathsf{D}_1, \mathsf{D}_2)$, $\mathsf{Y} = (\mathsf{Y}_1, \mathsf{Y}_2)$ and A contains all deterministic probability measures in $\Delta(D^2)$. If

$$\mathbb{P}_{\Box}^{\mathsf{Y}_1\mathsf{Y}_2|\mathsf{D}_1\mathsf{D}_2}(y_1, y_2 | d_1, d_2) = [\![(y_1, y_2) = (d_1 + d_2, d_1 + d_2)]\!] \tag{63}$$

Then $\mathbb{P}_{\delta_{00}}^{\mathsf{Y}_1\mathsf{D}_1}(y_1) = [\![y_1 = 0]\!]$ while $\mathbb{P}_{\delta_{01}}^{\mathsf{Y}_1} = [\![y_1 = 1]\!]$. However, $\delta_0 0^{\mathsf{D}_1} = \delta_{01}^{\mathsf{D}_1} = \delta_0^{\mathsf{D}_1}$ so $\mathbb{P}_{\Box}$ exhibits non-local consequences. However, taking $(d_i, d_j) := \delta_{d_i d_j} \in A$,

$$\mathbb{P}_{d_2, d_1}^{\mathsf{Y}_1\mathsf{D}_1\mathsf{Y}_2\mathsf{D}_2}(y_1, d_1, y_2, d_2) = [\![(y_1, y_2) = (d_2 + d_1, d_2 + d_1)]\!] \tag{64}$$

$$= [\![(y_2, y_1) = (d_1 + d_2, d_1 + d_2)]\!] \tag{65}$$

$$= \mathbb{P}_{d_1, d_2}^{\mathsf{Y}_1\mathsf{D}_1\mathsf{Y}_2\mathsf{D}_2}(y_2, d_2, y_1, d_1) \tag{66}$$

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

A conditional probability model that exhibits locality of consequences but does not commute with exchange:

Alternatively, suppose the same setup, but define $\mathbb{P}_\square$ instead by

$$\mathbb{P}_\square \mathsf{Y}_1\mathsf{Y}_2|\mathsf{D}_1\mathsf{D}_2(y_1,y_2|d_1,d_2) = [\![(y_1,y_2)=(0,1)]\!] \tag{67}$$

for all $\alpha \in A$.

Then $\mathbb{P}_\square$ exhibits locality of consequences. If $\mathbb{P}_\alpha^{\mathsf{D}_S} = \mathbb{P}_\beta^{\mathsf{D}_S}$ for $S \subset \{0,1\}$ then:

$$\begin{aligned}\mathbb{P}_\alpha^{\mathsf{Y}_S\mathsf{D}_S}(y_s,d_s) &= \sum_{y_2' \in \{0,1\}^{S^C}} [\![(y_1,y_2)=(0,1)]\!]\mathbb{P}_\alpha^{\mathsf{D}_S}(d_s) &(68)\\ &= \mathbb{P}_\beta^{\mathsf{Y}_S\mathsf{D}_S}(y_s,d_s) &(69)\end{aligned}$$

However, $\mathbb{P}_\square$ does not commute with exchange. For all $\alpha, \beta \in A$:

$$\begin{aligned}\mathbb{P}_\alpha \mathsf{Y}_1\mathsf{Y}_2(y_1,y_2) &= [\![(y_1,y_2)=(0,1)]\!] &(70)\\ &\neq \mathbb{P}_\beta \mathsf{Y}_1\mathsf{Y}_2(y_2,y_1) &(71)\end{aligned}$$

□

Although locality of consequences has a lot in common with an assumption non-interference, it still allows for some models in which exhibit certain kinds of interference between actions and outcomes of different indices. For example: I have an experiment where I first flip a coin and record the results of this flip as the outcome of the first step of the experiment, but I can choose either to record this same outcome as the provisional result of the second step (this is the choice $\mathsf{D}_1 = 0$), or choose to flip a second coin and record the result of that as the provisional result of the second step of the experiment (this is the choice $\mathsf{D}_1 = 1$). At the second step, I may further choose to copy the provisional results ($\mathsf{D}_2 = 0$) or invert them ($\mathsf{D}_2 = 1$). Then

- The marginal distribution of both experiments in isolation is Bernoulli(0.5) no matter what choices I make, so a model of this experiment would satisfies Definition 4.3
- Nevertheless, the choice for the first experiment affects the result of the second experiment

Note that this example would not satisfy exchange commutativity.

We call the conjunction of exchange commutativity and consequence locatlity *causal contractibility*.

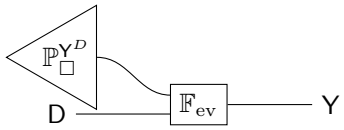
Definition 4.5 (Causal contractibility)**.** A conditional probability model $(\mathbb{P}_\square^{\overline{\mathsf{Y}|\mathsf{D}}}, A)$ is causally contractible if it is both commutative with exchange and commutative with marginalisation.

4.3 Repeatable consequence conditionals exist iff a model is causally contractible

The main result in this section is Theorem 4.7 which shows that a conditional probability model $\mathbb{P}_\Box$ is causally contractible if and only if it can be represented as the product of a distribution over hypotheses $\mathbb{P}_\Box^{\mathsf{H}}$ and a collection of identical conditional probabilities $\mathbb{P}_\Box^{\mathsf{Y}_1|\mathsf{D}_1\mathsf{H}}$. Note the hypothesis H that appears in this conditional; it can be given the interpretation of a random variable that expresses the "true but initially unknown" $\mathsf{Y}_1|\mathsf{D}_1$ conditional probability.

Lemma 4.6 (Exchangeable randomness pushback)**.** *A conditional probability model* $(\mathbb{P}_\Box^{\mathsf{Y}|\mathsf{D}}, A)$ *such that* $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ *and* $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$*.* $\mathbb{P}_\Box$ *is causally contractible if and only if there exists a column exchangeable probability distribution* $\mathbb{P}^{\mathsf{Y}^D}$ *such that*

$$\mathbb{P}_\Box^{\mathsf{Y}|\mathsf{D}} = \tag{72}$$



$$\iff \tag{73}$$

$$\mathbb{P}_\Box^{\mathsf{Y}|\mathsf{D}}(y|d) = \mathbb{P}^{(\mathsf{Y}^D_{d_i i})_{\mathbb{N}}}(y) \tag{74}$$

Where $\mathbb{F}_{ev}$ *is the Markov kernel associated with the evaluation map*

$$ev : D^{\mathbb{N}} \times Y^{D\times\mathbb{N}} \to Y \tag{75}$$

$$((d_i)_{\mathbb{N}}, (y_{ij})_{i\in D, j\in\mathbb{N}}) \mapsto (y_{d_i i})_{i\in\mathbb{N}} \tag{76}$$

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.

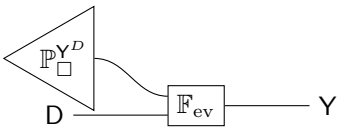
Define

$$\mathbb{P}^{\mathsf{Y}^D}((y_{ij})_{D\times\mathbb{N}}) := \mathbb{P}_e^{\mathsf{Y}}((y_{|D|i+j})_{i\in D, j\in\mathbb{N}}) \tag{77}$$

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 $i, j \in \mathbb{N}$.

$$\mathbb{Q} : D \nrightarrow Y \tag{78}$$

$$\mathbb{Q} := \tag{79}$$



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

$$\sum_{y\in\mathsf{Y}^{-1}(y_A)} \mathbb{Q}(y|d) = \sum_{y\in\mathsf{Y}^{-1}(y_A)} \mathbb{P}^{(\mathsf{Y}^D_{d_i i})_A}(y) \tag{80}$$

$$= \sum_{y\in\mathsf{Y}^{-1}(y_A)} \mathbb{P}_e^{(\mathsf{Y}_{|D|d_i+i})_A}(y) \tag{81}$$

$$= \mathbb{P}_e^{\mathsf{Y}_B}(y_A) \tag{82}$$

$$= \mathbb{P}_d^{\mathsf{Y}_A}(y_A) \qquad \text{by causal contractibility} \tag{83}$$

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

$$\mathbb{Q}(y|d) = \mathbb{P}_d^{\mathsf{Y}}(y) \tag{84}$$

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

$$\mathbb{Q}(y|d) = \mathbb{P}_d^{\mathsf{Y}|\mathsf{D}}(y|d) \tag{85}$$

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

$$\mathbb{P}^{\mathsf{Y}|\mathsf{D}} = \mathbb{Q} \tag{86}$$

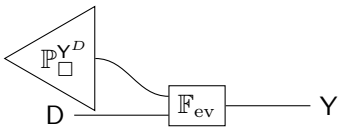
Next we will show $\mathbb{P}^{\mathsf{Y}^D}$ is contractible. Consider any subsequences Y^D_S and Y^D_T of Y^D with $|S| = |T|$. Let $\rho(S)$ be the "expansion" of the indices S, i.e. $\rho(S) = (|D|i+j)_{i\in S, j\in D}$. Then by construction of e, $e_{\rho(S)} = e_{\rho(T)}$ and therefore

$$\mathbb{P}^{\mathsf{Y}^D_S} = \mathbb{P}_e^{\mathsf{Y}_{\rho(S)}}) \tag{87}$$

$$= \mathbb{P}_e^{\mathsf{Y}_{\rho(T)}}) \quad \text{by contractibility of } \mathbb{P} \text{ and the equality } e_{\rho(S)} = e_{\rho(T)} \tag{88}$$

$$= \mathbb{P}^{\mathsf{Y}^D_T} \tag{89}$$

If: Suppose

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

(90)

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,

$$\mathbb{P}^{\mathsf{Y}_S|\mathsf{D}}(y_S|d) = \sum_{y_S^D \in Y^{|D|\times|S|}} \mathbb{P}^{\mathsf{Y}_S^D}(y_S^D)\mathbb{F}_{\text{ev}}(y_S|d, y_S^D) \tag{91}$$

$$= \sum_{y_S^D \in Y^{|D|\times|T|}} \mathbb{P}^{\mathsf{Y}_T^D}(y_S^D)\mathbb{F}_{\text{ev}}(y_S|d, y_S^D) \quad \text{by contractibility of } \mathbb{P}^{\mathsf{Y}_T^D} \tag{92}$$

$$= \mathbb{P}^{\mathsf{Y}_T|\mathsf{D}}(y_S|d) \tag{93}$$

□

As we pointed out, there are similarities between tabular distributions like $\mathbb{P}^{\mathsf{Y}^D}$ that appears in Lemma 4.6 and potential outcomes causal models. However, the $\mathbb{P}^{\mathsf{Y}^D}$ that appears in this lemma usually can't be interpreted as a distribution of potential outcomes. For example, consider a series of bets on fair coinflips. Model the consequence Y_i as uniform on $\{0,1\}$ for any decision D_i, for all i. Specifically, $D = Y = \{0,1\}$ 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 of $\mathbb{P}^{\mathsf{Y}^D}$ following the method in Lemma 4.6 yields $\mathbb{P}^{Y_i^D}(y_i^D) = \prod_{j\in D} 0.5$ for all $y_i^D \in Y^D$. In this model Y_i^0 and Y_i^1 are independent and uniformly distributed. However, if we wanted Y_i^0 to be interpretable as "what would happen if I bet on outcome 0 on turn i" and Y^1 to represent "what would happen if I bet on outcome 1 on turn i", then we ought to have $\mathsf{Y}_i^0 = 1 - \mathsf{Y}_i^1$.

Lemma 4.6 also does not establish that causal contractibility is necessary for the existence of a potential outcomes. A counterexample is any potential outcomes model with potential outcomes Z^D where the distribution $\mathbb{P}^{\mathsf{Z}^D}$ is not column exchangeable. Such a model is not causally contractible.

The tabular distribution $\mathbb{P}^{\mathsf{Y}^D}$ along with the evaluation function $\mathbb{F}_{\text{ev}}$ is a randomness pushback of the conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{D}}$. Because $\mathbb{P}^{\mathsf{Y}^D}$ is a column exchangeable probability distribution we can apply De Finetti's theorem to show $\mathbb{P}^{\mathsf{Y}^D}$ is representable as a product of identical parallel copies of $\mathbb{P}^{\mathsf{Y}_1^D|\mathsf{H}}$ and a common prior $\mathbb{P}^{\mathsf{H}}$. This in turn can be used to show that $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{D}}$ can be represented as a product of identical parallel copies of $\mathbb{P}_\square^{\mathsf{Y}_1|\mathsf{D}_1\mathsf{H}}$ and the same common prior $\mathbb{P}_\square^{\mathsf{H}}$. This is the main result: the copies of $\mathbb{P}_\square^{\mathsf{Y}_1|\mathsf{D}_1\mathsf{H}}$ are the repeatable response conditionals.

Theorem 4.7. *Suppose we have a sample space $(\Omega, \mathcal{F})$ and a conditional probability model $(\mathbb{P}_\square^{\mathsf{Y}|\mathsf{D}}, A)$ such that $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$. $\mathbb{P}_\square$ is causally contractible if and only if there exists some $\mathsf{H} : \Omega \to H$ such that $\mathbb{P}^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i}$ exists*

for all $i \in \mathbb{N}$ *and*

$$\mathbb{P}^{\mathsf{Y}|\mathsf{HD}} = \quad \mathsf{H} \quad \mathsf{D} \quad \boxed{\Pi_{D,i}} \quad \boxed{\mathbb{P}_{\square}^{\mathsf{Y}_0|\mathsf{HD}_0}} \quad \mathsf{Y}_i \quad i \in \mathbb{N} \tag{94}$$

$$\iff \tag{95}$$

$$\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}} \mathsf{Y}_{\mathbb{N}\setminus i}, \mathsf{D}_{\mathbb{N}\setminus i} | \mathsf{HD}_i \qquad \forall i \in \mathbb{N} \tag{96}$$

$$\wedge \mathbb{P}^{\mathsf{Y}_i|\mathsf{HD}_i} = \mathbb{P}^{\mathsf{Y}_0|\mathsf{HD}_0} \qquad \forall i \in \mathbb{N} \tag{97}$$

Where $\Pi_{D,i} : D^{\mathbb{N}} \twoheadrightarrow D$ *is the* i*th projection map.*

Proof. We make use of Lemma 4.6 to show that we can represent the conditional probability $\mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{D}}$ as

$$\mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{D}} = \quad \boxed{\mathbb{P}_{\square}^{\mathsf{Y}^D}} \quad \mathsf{D} \quad \boxed{\mathbb{F}_{\mathrm{ev}}} \quad \mathsf{Y} \tag{98}$$

$$\tag{99}$$

As a preliminary, we will show

$$\mathbb{F}_{\mathrm{ev}} = \quad \mathsf{H} \quad \mathsf{D} \quad \boxed{\Pi_{Y^D,i}} \quad \boxed{\Pi_{D,i}} \quad \boxed{\mathbb{F}_{\mathrm{ev},i}} \quad \mathsf{Y}_i \quad i \in \mathbb{N} \tag{100}$$

Where $\Pi_{Y^D,i} : Y^{D\times\mathbb{N}} \twoheadrightarrow Y^D$ is the ith column projection map on $\mathsf{Y}^{D\times\mathbb{N}}$ and $\mathrm{ev}_{Y^D\times D} : Y^D \times D \to Y$ is the evaluation function

$$((y_i)_{i\in D}, d) \mapsto y_d \tag{101}$$

Recall that ev is the function

$$((d_i)_{\mathbb{N}}, (y_{ij})_{i\in D, j\in\mathbb{N}}) \mapsto (y_{d_i i})_{i\in\mathbb{N}} \tag{102}$$

By definition, for any $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$

$$\mathbb{F}_{\mathrm{ev}}(\prod_{i\in\mathbb{N}} A_i | (d_i)_{\mathbb{N}}, (y_{ij})_{i\in D, j\in\mathbb{N}}) = \delta_{(y_{d_i i})_{i\in\mathbb{N}}}(\prod_{i\in\mathbb{N}} A_i) \tag{103}$$

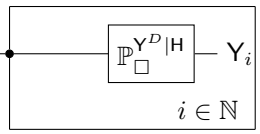
$$= \prod_{i\in\mathbb{N}} \delta_{y_{d_i i}}(A_i) \tag{104}$$

$$= \mathrm{copy}^{\mathbb{N}} \prod_{i\in\mathbb{N}} (\Pi_{D,i} \otimes \Pi_{Y,i}) \mathbb{F}_{\mathrm{ev}_{Y^D\times D}} \tag{105}$$

Which is what we wanted to show.

Only if: With $\mathbb{P}_{\square}^{\mathsf{Y}^D}$ column exchangeable. That is, letting $\mathsf{Y}^D = (\mathsf{Y}_i^D)_{i\in\mathbb{N}}$, the Y_i^D are exchangeable with respect to $\mathbb{P}_{\square}^{\mathsf{Y}^D}$. From kal (2005) we have a directing random measure H such that

$$\mathbb{P}_{\square}^{\mathsf{Y}^D|\mathsf{H}} = \tag{106}$$

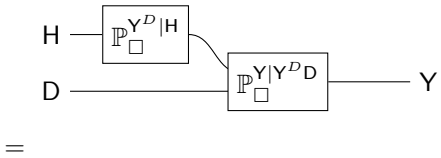


$$\iff \tag{107}$$

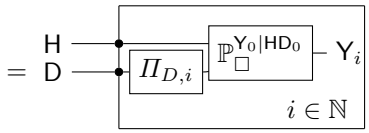
$$\mathbb{P}_{\square}^{\mathsf{Y}^D|\mathsf{H}}(\prod_{i\in\mathbb{N}} A_i|h) = \prod_{i\in\mathbb{N}} \mathbb{P}_{\square}^{\mathsf{Y}_0^D|\mathsf{H}}(A_i|h) \tag{108}$$

Furthermore, because Y is a deterministic function of D and Y^D, $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_{\square}} \mathsf{H}|(\mathsf{D}, \mathsf{Y}^D)$ and by definition of Y^D, $\mathsf{Y}^D \perp\!\!\!\perp_{\mathbb{P}_{\square}} \mathsf{D}$ and so

$$\mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{HD}} = \mathbb{P}_{\square}^{\mathsf{Y}^D|\mathsf{HD}} \odot \mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{Y}^D\mathsf{HD}} \tag{109}$$



$=$



(110)

If: By assumption

$$\mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{D}}(\prod_{i\in\mathbb{N}} A_i|h, (d_i)_{i\in\mathbb{N}}) = \int_H \prod_{i\in\mathbb{N}} \mathbb{P}_{\square}^{\mathsf{Y}_1|\mathsf{HD}_1}(A_i|h, d_i)\mathbb{P}_{\square}^{\mathsf{H}}(\mathrm{d}h) \tag{111}$$

Consider α, α' such that $\mathbb{P}_{\alpha}^{\mathsf{D}_M} = \mathbb{P}_{\alpha'}^{\mathsf{D}_L}$ for $L, M \subset \mathbb{N}$ with $|M| = |L|$, both finite. Then

$$\mathbb{P}_{\alpha}^{\mathsf{Y}_M}(A) = \int_{D^{\mathbb{N}}} \mathbb{P}_{\alpha}^{\mathsf{Y}_M|\mathsf{D}}(A|d)\mathbb{P}_{\alpha}^{\mathsf{D}}(\mathrm{d}d) \tag{112}$$

$$= \int_H \int_{D^{\mathbb{N}}} \prod_{i\in M} \mathbb{P}_{\square}^{\mathsf{Y}_1|\mathsf{HD}_1}(A_i|h,d_i)\mathbb{P}_{\alpha}^{\mathsf{D}}(\mathrm{d}d)\mathbb{P}_{\square}^{\mathsf{H}}(\mathrm{d}h) \tag{113}$$

$$= \int_H \int_{D^{|M|}} \prod_{i\in M} \mathbb{P}_{\square}^{\mathsf{Y}_1|\mathsf{HD}_1}(A_i|h,d_i)\mathbb{P}_{\alpha}^{\mathsf{D}_M}(\mathrm{d}d_M)\mathbb{P}_{\square}^{\mathsf{H}}(\mathrm{d}h) \tag{114}$$

$$= \int_H \int_{D^{|M|}} \prod_{i\in M} \mathbb{P}_{\square}^{\mathsf{Y}_1|\mathsf{HD}_1}(A_i|h,d_i)\mathbb{P}_{\alpha'}^{\mathsf{D}_N}(\mathrm{d}d_N)\mathbb{P}_{\square}^{\mathsf{H}}(\mathrm{d}h) \tag{115}$$

$$= \int_H \int_{D^{\mathbb{N}}} \prod_{i\in M} \mathbb{P}_{\square}^{\mathsf{Y}_1|\mathsf{HD}_1}(A_i|h,d_i)\mathbb{P}_{\alpha'}^{\mathsf{D}}(\mathrm{d}d)\mathbb{P}_{\square}^{\mathsf{H}}(\mathrm{d}h) \tag{116}$$

$$= \mathbb{P}_{\alpha'}^{\mathsf{Y}_M}(A) \tag{117}$$

□

4.4 Modelling different measurement procedures

An important question is: when is it reasonable to assume causal contractibility? We're going to focus just on the assumption of commutativity of exchange because we have more interesting things to say about it. There is a tempting but false line of argument one could adopt: $(\mathbb{P}_{\square}^{\mathsf{Y}_M|\mathsf{D}_M}, A)$ is a model of $|M|$ indistinguishable "experimental units", because they are indistinguishable they can be interchanged without altering the appropriate model, and so commutativity of exchange holds.

The problem with this line of reasoning is that interchangeability of "experimental units" doesn't imply commutativity of exchange. The problem is, roughly speaking, we may have indistinguishable experimental units when a decision function is chosen, but the decision function might leave some uncertainty over the actual decisions, which means the experimental units may be distinguishable when the actual decisions are made. If the decision function is deterministic, this possibility is ruled out. We'll explain this in more detail with an example, and in the next section we'll discuss randomisation.

4.5 Example: commutativity of exchange in the context of treatment choices

To justify an assumption of commutativity of exchange, we will argue as follows:

- Two measurement procedures should be considered equivalent in the sense that the same model is appropriate for both
- The models associated with the two procedures are related to one another by composition with the relevant swap maps

- Therefore the model associated with the first experiment is equivalent to the same model composed with the relevant swap maps

First, we want to spell out in detail how composing a model of one measurement procedure with a swap map can result in a model appliccable to a different measurement procedure. Recall that we assume that a single master measurement procedure $\mathcal{S}$ taking values in Ψ, and observables are all functions of $\mathcal{S}$. Given a model $(\mathbb{P}_\square, A)$ associated with $\mathcal{S}$, the model does not in general apply to an alternative measurement procedure $\mathcal{S}'$.

However, it is also a principle of measurement procedures that a measurement procedure followed by the application of a function is itself a measurement procedure. Thus a model $(\mathbb{P}_\square, A)$ associated with $\mathcal{S}$ may also be informative about a procedure $f \circ \mathcal{S}$ for any $f : \Psi \to X$.

In particular, consider measurement procedures related by *swaps*. For example, suppose we have $(\mathcal{D}_1, \mathcal{D}_2)$ and $(\mathcal{D}_1^{\text{swap}}, \mathcal{D}_2^{\text{swap}}) := (\mathcal{D}_2, \mathcal{D}_1)$. Then, given any probability model $\mathbb{P}_\alpha^{\mathsf{D}_1\mathsf{D}_2}$ we have $\mathbb{P}_\alpha^{\mathsf{D}_1^{\text{swap}}\mathsf{D}_2^{\text{swap}}} = \mathbb{P}_\alpha^{\mathsf{D}_1\mathsf{D}_2}$. In this way, $\mathbb{P}_\alpha^{\mathsf{D}_1\mathsf{D}_2}$ is a model of $(\mathcal{D}_1, \mathcal{D}_2)$ and induces a unique model of $(\mathcal{D}_1^{\text{swap}}, \mathcal{D}_2^{\text{swap}})$ via composition with a swap map.

Technically, this requires an assumption: if X is associated with $\mathcal{X}$ then $f \circ \mathsf{X}$ is associated with $f \circ \mathcal{X}$ (roughly: the abstract mathematical idea of composing a function with something and the actual process of applying a function to something and obtaining a result are treated as the same thing)

Concretely, commutativity of exchange can be justified if we suppose that the same model $(\mathbb{P}_\square^{\mathsf{Y}_M|\mathsf{D})_M}, A)$ should describe

- A measurement procedure $\mathcal{S}$ that yields $|M|$ outcomes $\mathcal{Y}_M$ and and $|M|$ decisions $\mathcal{D}_M$
- Any other $|M|$ outcomes $\mathcal{Y}_M^{\text{swap}}$ and $|M|$ decisions $\mathcal{D}_M^{\text{swap}}$, related to the originals by a swap.

Consider the following two scenarios:

1. Dr Alice is going to see two patients who are both complaining of lower back pain and are otherwise unknown to Alice. Prior to seeing them, she settles on a decision function α which deterministically sets her treatment choices according to a function decisions(α)
2. As before, but α is a "decision inclination" and $\mathbb{P}_\alpha^{\mathsf{D}_1\mathsf{D}_2}$ nondeterministic

Alice could model both situations with a sequential conditional probability model $(\mathbb{P}_\square^{\mathsf{Y}_1\mathsf{Y}_2|\mathsf{D}_1\mathsf{D}_2}, A)$ with the elements of A identified with probability models of the form $\mathbb{P}_\alpha^{\mathsf{D}_1\mathsf{D}_2}$. Might she, in one or both situations, consider this condiitonal probability model to be causally contractible?

We will assume that both satisfy commutativity of marginalisation – that is, the first patient's outcomes are expected to be the same no matter what is planned for the second patient and vise versa. We want to know if they satisfy commutativity of exchange.

The argument we want to make (if it can be supported) is:

- We can describe two measurement procedures that should share the same model
- The first is a measurement procedure for $(\mathsf{D}_1, \mathsf{D}_2, \mathsf{Y}_1, \mathsf{Y}_2)$
- The second is a measurement procedure for $(\mathsf{D}_1^{\text{swap}}, \mathsf{D}_2^{\text{swap}}, \mathsf{Y}_1^{\text{swap}^{-1}}, \mathsf{Y}_2^{\text{swap}^{-1}})$

At the outset, Alice does not know any features that might distinguish the two patients, so it is reasonable to think that she should adopt the same model for a) the original experiment and b) the same experiment, except with the patients interchanged. Note that interchanging *patients* does not correspond directly to any operation on the model $(\mathbb{P}_{\square}^{\mathsf{Y}_1\mathsf{Y}_2|\mathsf{D}_1\mathsf{D}_2}, A)$ which describes decisions and, not patients.

We will define measurement procedures using pseudocode, because we find it a lot easier to keep track of operations like swaps in this format. This presentation has the unintended effect of suggesting that measurement procedures are like computer programs. We're not sure if this is a helpful way to think about things – one of the key points of this example is that precise and imprecise measurement procedures may need quite different models, but thinking of measurement procedures as computer programs suggests that all measurement procedures are precise, which is not the case. Some steps may be precise, and we can express these steps with pseudocode, while other steps may be less precise.

Suppose the first scenario corresponds to the following procedure $\mathcal{S}$ which yields values in $A \times D^2 \times Y^2$. D_i is the projection $(\alpha, d_1, d_2, y_1, y_2) \mapsto d_i$ composed with $\mathcal{S}$ and Y_i is the projection $(\alpha, d_1, d_2, y_1, y_2) \mapsto y_i$ composed with $\mathcal{S}$.

procedure $\mathcal{S}$
  `assert`(patient A knowledge=patient B knowledge)
  $\alpha \leftarrow$ choose_alpha
  $(\mathcal{D}_1, \mathcal{D}_2) \leftarrow$ decisions(α)
  $\mathcal{Y}_1 \leftarrow$ apply($\mathcal{D}_1$, patient A)
  $\mathcal{Y}_2 \leftarrow$ apply($\mathcal{D}_2$, patient B)
  return $(\alpha, \mathcal{D}_1, \mathcal{D}_2, \mathcal{Y}_1, \mathcal{Y}_2)$
end procedure

Make the assumption that, on the basis that the patients are indistinguishable to Alice at the time of model construction, the same model is appropriate for the original measurement procedure and a modified measurement procedure in which the patients are swapped (we say the measurement procedures are "equivalent"). Assume also that swapping the order of treatment and swapping the order in which outcomes are recorded yields an equivalent measurment procedure (in Walley (1991)'s language, the first assumption is based on "symmetry of evidence" and the second on "evidence of symmetry"). Putting these two assumptions together, the following procedure $\mathcal{S}'$ is equivalent to the original:

procedure $\mathcal{S}'$
  `assert`(patient A knowledge=patient B knowledge)
  $\alpha \leftarrow$ choose_alpha
  $(\mathcal{D}_1, \mathcal{D}_2) \leftarrow$ decisions(α)

$\mathcal{Y}_2 \leftarrow \text{apply}(\mathcal{D}_2, \text{patient A})$
$\mathcal{Y}_1 \leftarrow \text{apply}(\mathcal{D}_1, \text{patient B})$
return $(\alpha, \mathcal{D}_1, \mathcal{D}_2, \mathcal{Y}_1, \mathcal{Y}_2)$
end procedure

Consider another measurement procedure $\mathcal{S}''$, which is a modified version of $\mathcal{S}$ where steps are added to swap decisions after they are chosen, then outcomes are swapped back once they have been observed:

procedure $\mathcal{S}''$
`assert`(patient A knowledge=patient B knowledge)
$\alpha \leftarrow$ choose_alpha
$(\mathcal{D}_1, \mathcal{D}_2) \leftarrow \text{decisions}(\alpha)$
$(\mathcal{D}_1^{\text{swap}}, \mathcal{D}_2^{\text{swap}}) \leftarrow (\mathcal{D}_2, \mathcal{D}_1)$
$\mathcal{Y}_1^{\text{swap}} \leftarrow \text{apply}(\mathcal{D}_1^{\text{swap}}, \text{patientA})$
$\mathcal{Y}_2^{\text{swap}} \leftarrow \text{apply}(\mathcal{D}_2^{\text{swap}}, \text{patient B})$
$(\mathcal{Y}_1, \mathcal{Y}_2) \leftarrow (\mathcal{Y}_2^{\text{swap}}, \mathcal{Y}_1^{\text{swap}})$
return $(\alpha, \mathcal{D}_1, \mathcal{D}_2, \mathcal{Y}_1, \mathcal{Y}_2)$
end procedure

Instead of explicitly performing the swaps, we can substitute $\mathcal{D}_2$ for $\mathcal{D}_1^{\text{swap}}$, $\mathcal{Y}_2$ for $\mathcal{Y}_1^{\text{swap}}$ and so on. The result is a procedure identical to $\mathcal{S}'$

procedure $\mathcal{S}''$
`assert`(patient A knowledge=patient B knowledge)
$\alpha \leftarrow$ choose_alpha
$(\mathcal{D}_1, \mathcal{D}_2) \leftarrow \text{decisions}(\alpha)$
$\mathcal{Y}_2 \leftarrow \text{apply}(\mathcal{D}_2, \text{patient A})$
$\mathcal{Y}_1 \leftarrow \text{apply}(\mathcal{D}_1, \text{patient B})$
return $(\alpha, \mathcal{D}_1, \mathcal{D}_2, \mathcal{Y}_1, \mathcal{Y}_2)$
end procedure

Thus $\mathcal{S}''$ is exactly the same as $\mathcal{S}'$, which by assumption is equivalent to the original $\mathcal{S}$, and so the assumptions of interchangeable patients and reversible order of treatment application imply the model should commute with exchange. Thus, if we could extend this example to an infinite sequence of patients, there would exist a Markov kernel $\mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{DH}} : D \times H \twoheadrightarrow Y$ representing a "definite but unknown causal consequence" shared by all experimental units.

This argument does *not* hold for scenario 2. In the absence of a deterministic function decisions(α) which defines the procedure for obtaining $\mathcal{D}_1$ and $\mathcal{D}_2$, there is some flexibility for how exactly these variables are measured (or chosen). In particular, we can posit measurement procedures such that permuting patients is not equivalent to permuting decisions and then appying the reverse permutation to outcomes.

For example, procedure $\mathcal{T}$ is compatible with scenario 2 (note that there are many procedures compatible with the given description)

procedure $\mathcal{T}$
`assert`(patient A knowledge=patient B knowledge)
$\alpha \leftarrow$ choose_alpha

patient A knowledge← inspect(patient A)
patient B knowledge← inspect(patient B)
$(\mathcal{D}_1, \mathcal{D}_2) \leftarrow$ vagueDecisions(α, patient A knowledge, patient B knowledge)
$\mathcal{Y}_1 \leftarrow$ apply($\mathcal{D}_1$, patient A)
$\mathcal{Y}_2 \leftarrow$ apply($\mathcal{D}_2$, patient B)
return $(\alpha, \mathcal{D}_1, \mathcal{D}_2, \mathcal{Y}_1, \mathcal{Y}_2)$
end procedure

Permutation of patients and treatment order now yields

procedure $\mathcal{T}'$
`assert`(patient A knowledge=patient B knowledge)
$\alpha \leftarrow$ choose_alpha
patient B knowledge← inspect(patient B)
patient A knowledge← inspect(patient A)
$(\mathcal{D}_1, \mathcal{D}_2) \leftarrow$ vagueDecisions(α, patient B knowledge, patient A knowledge)
$\mathcal{Y}_2 \leftarrow$ apply($\mathcal{D}_2$, patient A)
$\mathcal{Y}_1 \leftarrow$ apply($\mathcal{D}_1$, patient B)
return $(\alpha, \mathcal{D}_1, \mathcal{D}_2, \mathcal{Y}_1, \mathcal{Y}_2)$
end procedure

While paired permuation of decisions and outcomes yields

procedure $\mathcal{T}''$
`assert`(patient A knowledge=patient B knowledge)
$\alpha \leftarrow$ choose_alpha
patient A knowledge← inspect(patient A)
patient B knowledge← inspect(patient B)
$(\mathcal{D}_1, \mathcal{D}_2) \leftarrow$ vagueDecisions(α, patient A knowledge, patient B knowledge)
$\mathcal{Y}_2 \leftarrow$ apply($\mathcal{D}_2$, patient A)
$\mathcal{Y}_1 \leftarrow$ apply($\mathcal{D}_1$, patient B)
return $(\alpha, \mathcal{D}_1, \mathcal{D}_2, \mathcal{Y}_1, \mathcal{Y}_2)$
end procedure

$\mathcal{T}'$ is not the same as $\mathcal{T}''$. In scenario 1, because decisions were deterministic on α, there was no room to pick anything different once α was chosen, so it doesn't matter if we add patient inspection steps or not. In scenario 2, decisions are not deterministic and there is vagueness in the procedure, so it is possible to describe compatible procedures where decisions depend on patient characteristics, and this dependence is not "undone" by swapping decisions.

4.6 Causal consequences of non-deterministic variables

In the previous section we gave an example of how commutativity of exchange can hold when we have a sequence of decisions such that we accept the follwing:

- Reordering the time at which decisions are made is held to be of no consequence
- The available information relevant to each decision is symmetric at the time the decision function is adopted

- The decision function deterministically prescribes which decisions are taken

We also discussed how the absence of determinism undermines the argument for exchange commutativity.

The determinism assumption rules out choosing decisions randomly. However, if we have causal consequences for deterministic decision variables, it is sometimes possible to extend them to indeterministic variables.

Lemma 4.8. *Given* $(\mathbb{P}_\Box, A)$ *with decisions* D_M *and consequences* Y_M*, if* $\mathbb{P}_\Box^{\mathsf{Y}_M|\mathsf{D}_M}$ *is causally contractible with consequence map* $\mathbb{P}_\Box^{\mathsf{Y}_0|\mathsf{D}_0\mathsf{H}}$ *and there exists* $\mathsf{X}_i = f \circ \mathsf{Y}_i$ *for some* $f : Y \to X$ *such that* $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_\Box} \mathsf{D}_i | \mathsf{H}\mathsf{X}_i$ *for all* $i \in M$*, then a causally contractible conditional probability* $\mathbb{P}_\Box^{\mathsf{Y}_M|\mathsf{X}_M}$ *exists.*

Proof. We want to show $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_\Box} \mathsf{Y}_{\{i\}^C}\mathsf{X}_{\{i\}^C} | \mathsf{H}\mathsf{X}_i$ for all $i \in M$, $\mathbb{P}_\Box^{\mathsf{Y}_i|\mathsf{H}\mathsf{X}_i}$ exists for all $i \in M$ and $\mathbb{P}_\Box^{\mathsf{Y}_i|\mathsf{H}\mathsf{X}_i} = \mathbb{P}_\Box^{\mathsf{Y}_j|\mathsf{H}\mathsf{X}_j}$.

Because X_i is a function of Y_i, and $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_\Box} \mathsf{Y}_{\{i\}^C}\mathsf{D}_{\{i\}^C} | \mathsf{H}\mathsf{D}_i$, we also have $\mathsf{Y}\mathsf{X}_i \perp\!\!\!\perp_{\mathbb{P}_\Box} \mathsf{Y}_{\{i\}^C}\mathsf{X}_{\{i\}^C} | \mathsf{H}\mathsf{D}_i$, and by weak union $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_\Box} \mathsf{Y}_{\{i\}^C}\mathsf{X}_{\{i\}^C} | \mathsf{H}\mathsf{D}_i\mathsf{X}_i$

Thus by contraction, $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_\Box} \mathsf{Y}_{\{i\}^C}\mathsf{D}_M | \mathsf{H}\mathsf{X}_i$.

By Corollary 5.22 and the existence of $\mathbb{P}^{\mathsf{Y}_i\mathsf{X}_i|\mathsf{H}\mathsf{D}_i}$ for all $i \in M$, $\mathbb{P}_\Box^{\mathsf{Y}_i|\mathsf{H}\mathsf{X}_i}$ exists for all i. Furthermore, because $\mathbb{P}^{\mathsf{Y}_i\mathsf{X}_i|\mathsf{H}\mathsf{D}_i} = \mathbb{P}^{\mathsf{Y}_j\mathsf{X}_j|\mathsf{H}\mathsf{D}_j}$ for all $i, j \in M$, $\mathbb{P}_\Box^{\mathsf{Y}_i|\mathsf{H}\mathsf{X}_i} = \mathbb{P}_\Box^{\mathsf{Y}_j|\mathsf{H}\mathsf{X}_j}$ for all $i, j \in M$. □

If the condition $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_\Box} \mathsf{D}_i | \mathsf{H}\mathsf{X}_i$ for all $i \in M$, we can say X_i is a proxy for controlling Y_i.

As an example of this, suppose $\mathsf{X} : \Omega \to X$ is a source of random numbers, the set of decisions D is a set of functions $X \to T$ for treatments $\mathsf{T} : \Omega \to T$ and $\mathsf{W} : \Omega \to W$ are the ultimate patient outcomes, with $\mathsf{Y}_i = (\mathsf{W}_i, \mathsf{T}_i)$. Then it may be reasonable to assume that $\mathsf{W}_i \perp\!\!\!\perp (\mathsf{D}_i, \mathsf{X}_i) | \mathsf{T}_i\mathsf{H}$ (where conditioning on H can be thought of as saying that this independence holds under infinite sample size). In this case, T_i is a proxy for controlling Y_i, and there exists a causal consequence $\mathbb{P}_\Box^{\mathsf{Y}_0|\mathsf{T}_0\mathsf{H}}$.

A "causal consequence of body mass index" is unlikely to exist on the basis of symmetric information and deterministic decisions because there are no actions available to set body mass index deterministically. However, given an underlying problem where we have symmetric information over a collection of patients and some kind of decision that can be made deterministically, causal consequences of body mass index may exist if body mass index is a proxy for controlling the outcomes of interest.

4.7 Intersubjective causal consequences

While the assumption of causal contractibility itself does not depend on any notion of subjectivity, our discussion of the applicability of this assumption assumed that a conditional probability model was being used to model Dr Alice's subjective uncertain knowledge. Crucially, the justification hinged on an assumption of the symmetry of Alice's information regarding different patients.

Causal inference is often performed in an intersubjective setting, where Ben might perform the experimeng, Carmel might do the analysis and Dr Alice make the ultimate decisions. This complicates the question of when the assumption of causal contractibility is appliccable. We leave the appropriate way to generalise this theory to such a setting open.

5 Appendix, needs to be organised

5.1 Existence of conditional probabilities

Lemma 5.1 (Conditional pushforward)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_{\{\}}$ with conditional $\mathbb{P}_{\{\}}^{\mathsf{X}|\mathsf{Y}}$ such that $\mathsf{Z} = f \circ \mathsf{Y}$ for some $f : Y \to Z$. Then there exists a conditional probability $\mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{X}} = \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \mathbb{F}_f$.*

Proof. Note that $(\mathsf{X}, \mathsf{Z}) = (\mathrm{id}_X \otimes f) \circ (\mathsf{X}, \mathsf{Y})$. Thus, by Lemma 2.8, for any $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$

$$\mathbb{P}_\alpha^{\mathsf{X}\mathsf{Z}} = \mathbb{P}_\alpha^{\mathsf{X}\mathsf{Y}} \mathbb{F}_{\mathrm{id}_X \otimes f} \tag{118}$$

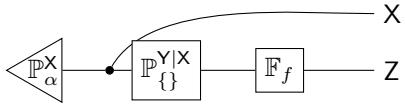
Note also that for all $A \in \mathcal{X}$, $B \in \mathcal{Z}$, $x \in X$, $y \in Y$:

$$\begin{aligned} \mathbb{F}_{\mathrm{id}_X \otimes f}(A \times B | x, y) &= \delta_x(A) \delta_{f(y)}(B) && (119)\\ &= \mathbb{F}_{\mathrm{id}_X}(A|x) \otimes \mathbb{F}_f(B|y) && (120)\\ \implies \mathbb{F}_{\mathrm{id}_X \otimes f} &= \mathbb{F}_{\mathrm{id}_X} \otimes \mathbb{F}_f && (121) \end{aligned}$$

Thus

$$\mathbb{P}_\alpha^{\mathsf{X}\mathsf{Z}} = (\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}) \mathbb{F}_{\mathrm{id}_X} \otimes \mathbb{F}_f \tag{122}$$

$$= \tag{123}$$



Which implies $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \mathbb{F}_f$ is a version of $\mathbb{P}_\alpha^{\mathsf{Z}|\mathsf{X}}$. Because this holds for all α, it is therefore also a version of $\mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{X}}$. □

Theorem 5.2 (Existence of regular conditionals)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$ with Y standard measurable and a probability model $\mathbb{P}_\alpha$ on $(\Omega, \mathcal{F})$. Then there exists a conditional $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}$.*

Proof. This is a standard result, see for example Çinlar (2011) Theorem 2.18. □

Theorem 5.3 (Existence of higher order valid conditionals with respect to probability sets)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_{\{\}}$ with regular conditional $\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$ and Y and Z standard measurable. Then there exists a regular $\mathbb{P}_{\{\}}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$.*

Proof. Given a Borel measurable map $m : X \nrightarrow Y \times Z$ let $f : Y \times Z \to Y$ be the projection onto Y. Then $f \circ (\mathsf{Y}, \mathsf{Z}) = \mathsf{Y}$. Bogachev and Malofeev (2020), Theorem 3.5 proves that there exists a Borel measurable map $n : X \times Y \nrightarrow Y \times Z$ such that

$$n(f^{-1}(y)|x, y) = 1 \tag{124}$$

$$m(\mathsf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x, y) m\mathbb{F}_f(dy|x) \forall A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z} \tag{125}$$

In particular, $\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$ is a Borel measurable map $X \nrightarrow Y \times Z$. Thus equation 125 implies for all $A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z}$

$$\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}(\mathsf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x, y) \mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}} \mathbb{F}_f(dy|x) \tag{126}$$

$$= \int_A n(B|x, y) \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}(dy|x) \tag{127}$$

Where Equation 127 follows from Lemma 5.1.
Then, for any $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$

$$\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}(\mathsf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x, y) \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}(dy|x) \tag{128}$$

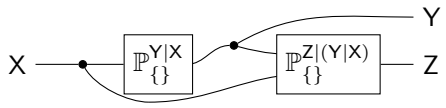
which implies n is a version of $\mathbb{P}_{\{\}}^{\mathsf{YZ}|(\mathsf{Y}|\mathsf{X})}$. By Lemma 5.1, $n\mathbb{F}_f$ is a version of $\mathbb{P}_{\{\}}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$. □

We might be motivated to ask whether the higher order conditionals in Theorem 5.3 can be chosen to be valid. Despite Lemma 5.8 showing that the existence of proper conditional probabilities implies the existence of valid ones, we cannot make use of this in the above theorem because Equation 124 makes n proper with respect to the "wrong" sample space $(Y \times Z, \mathcal{Y} \otimes \mathcal{Z})$ while what we would need is a proper conditional probability with respect to $(\Omega, \mathcal{F})$.

We can choose higher order conditionals to be valid in the case of discrete sets, and whether we can choose them to be valid in more general measurable spaces is an open question.

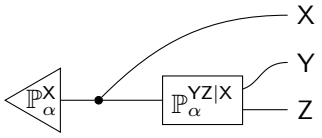
Theorem 5.4 (Higher order conditionals)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_{\{\}}$ with conditional $\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$. Then $\mathbb{P}_{\{\}}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ is a version of $\mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}$*

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

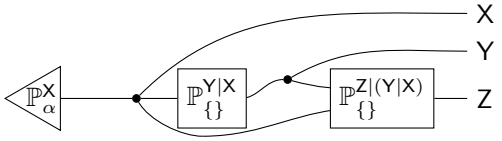


$$\mathbb{P}^{\mathsf{YZ}|\mathsf{X}}_{\alpha} = \tag{129}$$

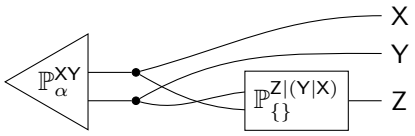
$$\implies \mathbb{P}^{\mathsf{XYZ}}_{\alpha} = \tag{130}$$



$$= \tag{131}$$



$$= \tag{132}$$

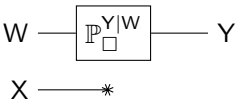


Thus $\mathbb{P}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}_{\{\}}$ is a version of $\mathbb{P}^{\mathsf{Z}|\mathsf{YX}}_{\alpha}$ for all α and hence also a version of $\mathbb{P}^{\mathsf{Z}|\mathsf{YX}}_{\{\}}$. □

Theorem 5.5. *Given probability gap model* $\mathbb{P}_{\{\}}$, X, Y, Z *such that* $\mathbb{P}^{\mathsf{Z}|\mathsf{YX}}_{\{\}}$ *exists,* $\mathbb{P}^{\mathsf{Z}|\mathsf{Y}}_{\{\}}$ *exists iff* $\mathsf{Z} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{X}|\mathsf{Y}$.

Proof. If: If $\mathsf{Z} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{X}|\mathsf{Y}$ then by Theorem 5.15, for each $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$ there exists $\mathbb{P}^{\mathsf{Z}|\mathsf{Y}}_{\alpha}$ such that

$$\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}_{\alpha} = \tag{133}$$



□

Theorem 5.6 (Valid higher order conditionals)**.** *Suppose we have a sample space* $(\Omega, \mathcal{F})$*, variables* $\mathsf{X} : \Omega \to X$ *and* $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ *and a probability set* $\mathbb{P}_{\{\}}$ *with regular conditional* $\mathbb{P}^{\mathsf{YZ}|\mathsf{X}}_{\{\}}$*, Y discrete and Z standard measurable. Then there exists a valid regular* $\mathbb{P}^{\mathsf{Z}|\mathsf{XY}}_{\{\}}$.

Proof. By Theorem 5.3, we have a higher order conditional $\mathbb{P}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}_{\{\}}$ which, by Theorem 5.4 is also a version of $\mathbb{P}^{\mathsf{Z}|\mathsf{XY}}_{\{\}}$.

We will show that there is a Markov kernel $\mathbb{Q}$ almost surely equal to $\mathbb{P}^{\mathsf{Z}|\mathsf{XY}}_{\{\}}$ which is also valid. For all $x, y \in X \times Y$, $A \in \mathcal{Z}$ such that $(\mathsf{X}, \mathsf{Y}, \mathsf{Z}) \bowtie \{(x, y)\} \times A = \emptyset$, let $\mathbb{Q}(A|x, y) = \mathbb{P}^{\mathsf{Z}|\mathsf{XY}}_{\{\}}(A|x, y)$.

By validity of $\mathbb{P}^{\mathsf{YZ}|\mathsf{X}}_{\{\}}$, $x \in \mathsf{X}(\Omega)$ and $(\mathsf{X}, \mathsf{Y}, \mathsf{Z}) \bowtie \{(x, y)\} \times A = \emptyset$ implies $\mathbb{P}^{\mathsf{YZ}|\mathsf{X}}_{\{\}}(\{y\} \times A|x) = 0$. Thus we need to show

$$\forall A \in \mathcal{Z}, x \in X, y \in Y : \mathbb{P}^{\mathsf{YZ}|\mathsf{X}}_{\{\}}(\{y\} \times A|x) = 0 \implies (\mathbb{Q}(A|x, y) = 0) \vee ((\mathsf{X}, \mathsf{Y}) \bowtie \{(x, y)\} = \emptyset) \tag{134}$$

For all x, y such that $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\{\}}(\{y\}|x)$ is positive, we have $\mathbb{P}^{\mathsf{YZ}|\mathsf{X}}(\{y\} \times A|x) = 0 \implies \mathbb{P}^{\mathsf{Z}|\mathsf{XY}}_{\square}(A|x, y) = 0 =: \mathbb{Q}(A|x, y)$.

Furthermore, where $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\{\}}(\{y\}|x) = 0$, we either have $(\mathsf{X}, \mathsf{Y}, \mathsf{Z}) \bowtie \{(x, y)\} \times A = \emptyset$ or can choose some $\omega \in (\mathsf{X}, \mathsf{Y}, \mathsf{Z}) \bowtie \{(x, y)\} \times A$ and let $\mathbb{Q}(\mathsf{Z}(\omega)|x, y) = 1$. This is an arbitrary choice, and may differ from the original $\mathbb{P}^{\mathsf{Z}|\mathsf{XY}}_{\{\}}$. However, because Y is discrete the union of all points y where $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\{\}}(\{y\}|x) = 0$ is a measure zero set, and so $\mathbb{Q}$ differs from $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\{\}}$ on a measure zero set. □

5.2 Validity

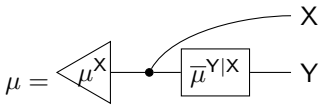
Validity is related to *proper* conditional probabilities. In particular, valid conditional probabilities exist when regular proper conditional probabilities exist.

Definition 5.7 (Regular proper conditional probability)**.** Given a probability space $(\mu, \Omega, \mathcal{F})$ and a variable $\mathsf{X} : \Omega \to X$, a regular proper conditional probbability $\mu^{|\mathsf{X}} : X \to\!\!\!\!\bullet\ \Omega$ is Markov kernel such that

$$\mu(A \cap \mathsf{X}^{-1}(B)) = \int_B \mu^{|\mathsf{X}}(A|x)\mu^{\mathsf{X}}(\mathrm{d}x) \qquad \forall A \in \mathcal{X}, B \in \mathcal{F} \tag{135}$$

$$\iff \tag{136}$$

$$\mu = \tag{137}$$



and

$$\mu^{|\mathsf{X}}(\mathsf{X}^{-1}(A)|x) = \delta_x(A) \tag{138}$$

Lemma 5.8. *Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, if there is a regular proper conditional probability $\mu^{|\mathsf{X}} : X \to\!\!\!\!\bullet\ \Omega$ then there is a valid conditional distribution $\mu^{\mathsf{Y}|\mathsf{X}}$.*

Proof. Take $\mathbb{K} = \mu^{|\mathsf{X}}\mathbb{F}_{\mathsf{Y}}$. We will show that $\mathbb{K}$ is valid, and a version of $\mu^{\mathsf{Y}|\mathsf{X}}$.

Defining $\mathsf{O} := \mathrm{id}_\Omega$ (the identity function $\Omega \to \Omega$), $\mu^{\mathsf{I}\mathsf{X}}$ is a version of $\mu^{\mathsf{O}|\mathsf{X}}$. Note also that $\mathsf{Y} = \mathsf{Y} \circ \mathsf{O}$. Thus by Lemma 5.1, $\mathbb{K}$ is a version of $\mu^{\mathsf{Y}|\mathsf{X}}$.

It remains to be shown that $\mathbb{K}$ is valid. Consider some $x \in X$, $A \in \mathcal{Y}$ such that $\mathsf{X}^{-1}(\{x\}) \cap \mathsf{Y}^{-1}(A) = \emptyset$. Then by the assumption $\mu^{\mathsf{I}\mathsf{X}}$ is proper

$$\mathbb{K}(\mathsf{Y} \bowtie A|x) = \delta_x(\mathsf{Y}^{-1}(A)) \tag{139}$$
$$= 0 \tag{140}$$

Thus $\mathbb{K}$ is valid. □

Theorem 5.9 (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

$$\mu^{\mathsf{X}}(x) = \mu(\mathsf{X} \bowtie x) \tag{141}$$
$$= \sum_{x' \in X} \mathbb{J}(x')\mathbb{K}(\mathsf{X} \bowtie x|x') \tag{142}$$
$$= 0 \tag{143}$$
$$\neq \mathbb{J}(x) \tag{144}$$

□

Lemma 5.10 (Semidirect product defines 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 maximal probability sets $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}[M]}$ and $\mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{Y}\mathsf{X}[M]}$ then defining*

$$\mathbb{R}_{\{\}}^{\mathsf{Y}\mathsf{Z}|\mathsf{X}} := \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{Y}\mathsf{X}} \tag{145}$$

we have

$$\mathbb{R}_{\{\}}^{\mathsf{Y}\mathsf{Z}|\mathsf{X}[M]} = \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}[M]} \cap \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{Y}\mathsf{X}[M]} \tag{146}$$

Proof. For any $\mathbb{R}_a \in \mathbb{R}_{\{\}}$

$$\mathbb{R}_a^{\mathsf{X}\mathsf{Y}\mathsf{Z}} = \mathbb{R}_a^{\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{Y}\mathsf{X}} \tag{147}$$
$$\implies \mathbb{R}_a^{\mathsf{X}\mathsf{Y}} = \mathbb{R}_a^{\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \tag{148}$$
$$\wedge \mathbb{R}_a^{\mathsf{X}\mathsf{Y}\mathsf{Z}} = \mathbb{R}_a^{\mathsf{X}\mathsf{Y}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{Y}\mathsf{X}} \tag{149}$$

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

Lemma 5.11 (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

$$\forall A \in \mathcal{X} : \mathsf{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathsf{X}|*}(A|*) = 0 \tag{150}$$

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

$$\forall A \in \mathcal{X} : \mathsf{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathsf{X}}(A) = 0 \tag{151}$$

□

Lemma 5.12 (Semidirect 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 $\mathbb{P}^{\mathsf{Y|X}}$ and $\mathbb{Q}^{\mathsf{Z|YX}}$, $\mathbb{P}^{\mathsf{Y|X}} \odot \mathbb{Q}^{\mathsf{Z|YX}}$ is also a valid candidate conditional.*

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

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

$$\begin{aligned} \mathbb{R}^{\mathsf{YZ|X}}(B \times C|x) &= \int_B \mathbb{Q}^{\mathsf{Z|YX}}(C|y,x)\mathbb{P}^{\mathsf{Y|X}}(dy|x) && (152)\\ &\leq \mathbb{P}^{\mathsf{Y|X}}(B|x) && (153)\\ &= 0 && (154) \end{aligned}$$

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

$$\begin{aligned} \mathbb{R}^{\mathsf{YZ|X}}(B \times C|x) &= \int_B \mathbb{Q}^{\mathsf{Z|YX}}(C|y,x)\mathbb{P}^{\mathsf{Y|X}}(dy|x) && (155)\\ &= 0 && (156) \end{aligned}$$

□

Corollary 5.13 (Valid conditionals are validly extendable to valid distributions)**.** *Given Ω, $\mathsf{U} : \Omega \to U$, $\mathsf{W} : \Omega \to W$ and a valid conditional $\mathbb{T}^{\mathsf{W}|\mathsf{U}}$, then for any valid conditional $\mathbb{V}^{\mathsf{U}}$, $\mathbb{V}^{\mathsf{U}} \odot \mathbb{T}^{\mathsf{W}|\mathsf{U}}$ is a valid probability.*

Proof. Applying Lemma 5.12 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 5.11. □

Theorem 5.14 (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 5.13.

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

$$(\mathbb{V}^{\mathsf{X}} \odot \mathbb{T}^{\mathsf{Y}|\mathsf{X}})(x, y) = \mathbb{T}^{\mathsf{Y}|\mathsf{X}}(y|x)\mathbb{V}^{\mathsf{X}}(x) \tag{157}$$

$$= \mathbb{T}^{\mathsf{Y}|\mathsf{X}}(y|x) \tag{158}$$

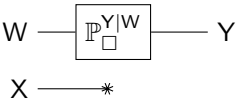
$$> 0 \tag{159}$$

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

5.3 Conditional independence

Theorem 5.15. *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*

$$\mathbb{P}^{\mathsf{Y}|\mathsf{WX}} = \tag{160}$$



$$\iff \mathbb{P}^{\mathsf{Y}|\mathsf{WX}}(y|w, x) = \mathbb{P}^{\mathsf{Y}|\mathsf{W}}(y|w) \tag{161}$$

Proof. See Cho and Jacobs (2019). □

5.4 Extended conditional independence

Constantinou and Dawid (2017) introduced the idea of *extended conditional independence*, which is a notion of conditional independence with respect to a parametrised collection of probability measures. It is motivated in part by the observation that such parametrised collections can be used to model causal questions. Furthermore, probability sets are closely related to parametrised probability sets – one can get the former from the latter by simply dropping the parameters.

This needs major revision, and is not a top priority right now

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 , 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:

Definition 5.16 (Extended conditional independence)**.** adf

Theorem 5.17. *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* $\mathbb{P}_{\alpha}^{\mathsf{VW}} := \mathbb{P}_{\alpha}^{\mathsf{W}} \odot \mathbb{P}^{\mathsf{V}|\mathsf{W}}$

$$\mathsf{A} \perp\!\!\!\perp_{\mathbb{P},ext} (\mathsf{B}, \mathsf{D}) | (\mathsf{C}, \mathsf{E}) \iff \mathsf{A}^* \perp\!\!\!\perp_{\mathbb{P}_\alpha} (\mathsf{B}^*, \mathsf{D}^*) | (\mathsf{C}^*, \mathsf{E}^*) \tag{162}$$

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

Theorem 5.18. *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* $\mathbb{P}_{\alpha}^{\mathsf{VW}} := \mathbb{P}_{\alpha}^{\mathsf{W}} \odot \mathbb{P}^{\mathsf{V}|\mathsf{W}}$

$$\mathsf{A} \perp\!\!\!\perp^{1}_{\mathbb{P},ext} (\mathsf{B}, \mathsf{D}) | (\mathsf{C}, \mathsf{E}) \iff \mathsf{A}^* \perp\!\!\!\perp_{\mathbb{P}} (\mathsf{B}^*, \mathsf{D}^*) | (\mathsf{C}^*, \mathsf{E}^*) \tag{163}$$

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

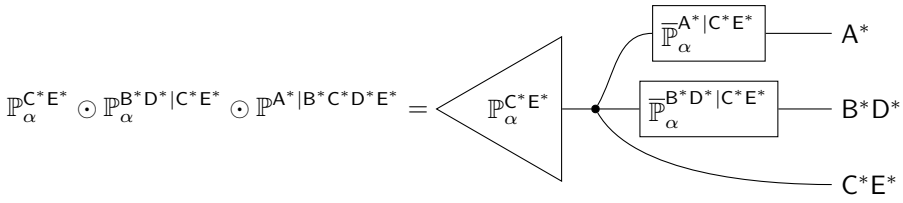
Only if:

For any β, $\mathbb{P}_{\beta}^{\mathsf{ABC}|\mathsf{DE}} = \mathbb{P}_{\beta}^{\mathsf{DE}} \odot \mathbb{P}^{\mathsf{ABC}|\mathsf{DE}}$. By Lemma 5.4, we have $\mathbb{P}^{\mathsf{A}|\mathsf{BCDE}}$ such that

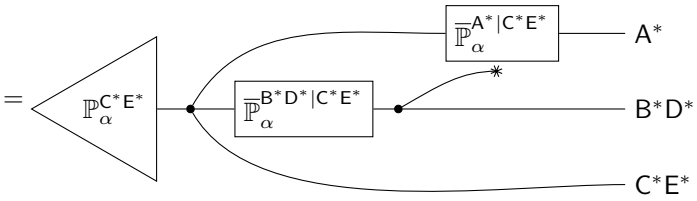
$$\begin{aligned}
\mathbb{P}_{\beta}^{\mathsf{A}^*\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} &= \mathbb{P}_{\beta}^{\mathsf{D}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{B}^*\mathsf{C}^*|\mathsf{D}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{A}^*|\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} && (164)\\
&= \mathbb{P}_{\beta}^{\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{A}^*|\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} && (165)\\
&= \mathbb{P}_{\beta}^{\mathsf{C}^*\mathsf{E}^*} \odot \mathbb{P}_{\beta}^{\mathsf{B}^*\mathsf{D}^*|\mathsf{C}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{A}^*|\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} && (166)
\end{aligned}$$

By Theorem 5.17, 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 $\mathbb{P}^{\mathsf{A}|\mathsf{BCDE}}$:



(167)

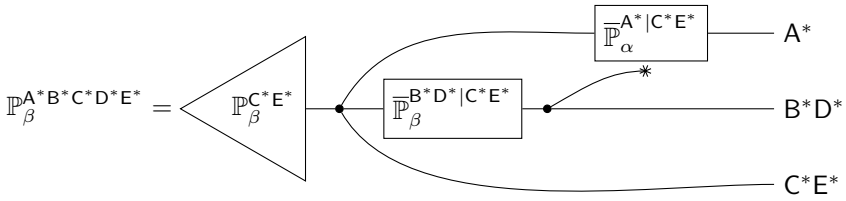


(168)

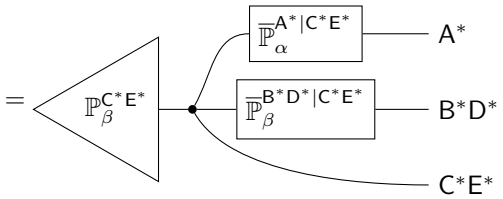
$$= \mathbb{P}_\alpha^{C^*E^*} \odot \mathbb{P}_\alpha^{B^*D^*|C^*E^*} \odot (\mathbb{P}_\alpha^{A^*|C^*E^*} \otimes \mathrm{erase}_{BD})$$

(169)

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



(170)



(171)

□

Conditional independence is a property of variables, we define "unresponsiveness" as a property of Markov kernels.

Definition 5.19 (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

$$\mathbb{P}_\square^{\mathsf{Y}|\mathsf{WX}} = \quad \mathsf{W} \text{ — } \boxed{\mathbb{P}_\square^{\mathsf{Y}|\mathsf{W}}} \text{ — } \mathsf{Y} \qquad \mathsf{X} \text{ ——} * \tag{172}$$

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

Definition 5.20 (Domination)**.** Given a probability set $\mathbb{P}_{\{\}} \subset \Delta(\Omega)$, $\mathbb{P}_\alpha$ dominates $\mathbb{P}_{\{\}}$ if $\mathbb{P}_\beta(B) > 0 \implies \mathbb{P}_\alpha(B) > 0$ for all $\mathbb{P}_\beta in \mathbb{P}_{\{\}}$, $B \in \mathcal{F}$.

Theorem 5.21 (Conditional independence from kernel unresponsiveness)**.** *Given standard measurable Ω, variables $\mathsf{W} : \Omega \to W$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and a probability set $\mathbb{P}_{\{\}} : A \to \Delta(\Omega)$ with conditional probability $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{WX}}$ such that there is $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$ dominating $\mathbb{P}_{\{\}}$, $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{X}|\mathsf{W}$ if and only if there is a version of $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{WX}}$ unresponsive to W.*

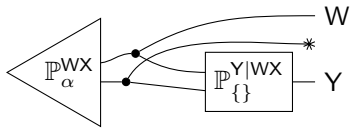
Proof. If: For every $\alpha \in A$ we can write

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}} = \quad \mathsf{W} \text{ — } \boxed{\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{W}}} \text{ — } \mathsf{Y} \qquad \mathsf{X} \text{ ——} * \tag{173}$$

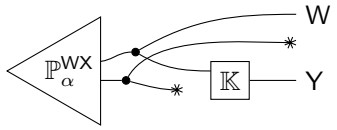
And so, by Theorem 5.15, $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{X}|\mathsf{W}$ for all $\alpha \in A$, and so $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{X}|\mathsf{W}$. Only if: For $\mathbb{P}_\alpha$ dominating $\mathbb{P}_{\{\}}$, by Theorem 5.15, there exists a version of $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}}$ unresponsive to W. Because $\mathbb{P}_\alpha$ dominates $\mathbb{P}_{\{\}}$, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}}$ differs from $\mathbb{P}_\beta^{\mathsf{Y}|\mathsf{WX}}$ on a set of measure 0 for any $\mathbb{P}_\beta \in \mathbb{P}_{\{\}}$, thus $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}}$ is a version of $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{WX}}$ also. □

Corollary 5.22. *Given standard measurable Ω, variables $\mathsf{W} : \Omega \to W$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and a probability set $\mathbb{P}_{\{\}} : A \to \Delta(\Omega)$ with conditional probability $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{WX}}$, $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{W}}$ exists if $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{X}|\mathsf{W}$.*

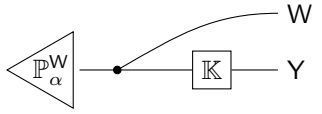
Proof. By Theorem 5.21, there is $\mathbb{K} : W \to\!\!\!\!\bullet\; Y$ such that for all α



$$\mathbb{P}_\alpha^{\mathsf{WY}} = \tag{174}$$



$$= \tag{175}$$



$$= \tag{176}$$

Thus $\mathbb{K}$ is a version of $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{W}}$. □

This result can fail to hold in the absence of 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.

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

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.

A. V. Banerjee, S. Chassang, and E. Snowberg. Chapter 4 - Decision Theoretic Approaches to Experiment Design and External ValidityaaWe thank Esther Duflo for her leadership on the handbookăand for extensive comments on earlier drafts. Chassang and Snowberg gratefully acknowledge the support of NSF grant SES-1156154. In Abhijit Vinayak Banerjee and Esther Duflo, editors, *Handbook of Economic Field Experiments*, volume 1 of *Handbook of Field Experiments*, pages 141–174. North-Holland, January 2017. doi: 10.1016/bs.hefe.2016.08.005. URL https://www.sciencedirect.com/science/article/pii/S2214658X16300071.

Abhijit V. Banerjee, James Berry, Esther Duflo, Harini Kannan, and Shobhini Mukerji. Mainstreaming an Effective Intervention: Evidence from Randomized Eavluations of 'Teaching at the Right Level' in India. SSRN Scholarly Paper ID 2843569, Social Science Research Network, Rochester, NY, September 2016. URL https://papers.ssrn.com/abstract=2843569.

Vladimir Bogachev and Ilya Malofeev. Kantorovich problems and conditional measures depending on a parameter. *Journal of Mathematical Analysis and Applications*, 486:123883, June 2020. doi: 10.1016/j.jmaa.2020.123883.

Erhan Çinlar. *Probability and Stochastics*. Springer, 2011.

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.

Bruno de Finetti. Foresight: Its Logical Laws, Its Subjective Sources. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Foundations and Basic Theory*, Springer Series in Statistics, pages 134–174. Springer, New York, NY, [1937] 1992. ISBN 978-1-4612-0919-5. doi: 10.1007/978-1-4612-0919-5_10. URL https://doi.org/10.1007/978-1-4612-0919-5_10.

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.

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

SANDER GREENLAND and JAMES M ROBINS. Identifiability, Exchangeability, and Epidemiological Confounding. *International Journal of Epidemiology*, 15(3):413–419, September 1986. ISSN 0300-5771. doi: 10.1093/ije/15.3.413. URL https://doi.org/10.1093/ije/15.3.413.

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

M. A. Hernán and S. L. Taubman. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity*, 32(S3):S8–S14, August 2008. ISSN 1476-5497. doi: 10.1038/ijo.2008.82. URL https://www.nature.com/articles/ijo200882.

KyleW (https://math.stackexchange.com/users/252356/kylew). Distribution of universal quantifiers over implication. Mathematics Stack Exchange. URL https://math.stackexchange.com/q/1377555. URL:https://math.stackexchange.com/q/1377555 (version: 2015-07-29).

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.

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.

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.

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.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.

Appendix: