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

David Johnston

January 21, 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 $\mathsf{Y}_i \perp\!\!\!\perp \mathsf{D}_i | \mathsf{H}\mathsf{X}_i$

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

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

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

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

2.1.1 Brief outline of probability gap models

We consider a probability model to be a probability space $(\Omega, \mathcal{F}, \mu)$ along with a collection of random variables. However, if I want to use probabilistic models to support decision making, then I need function from options to probability models. For example, suppose I have two options $A = \{0, 1\}$, and I want to compare these options based on what I expect to happen if I choose them. If I choose option 0, then I can (perhaps) represent my expectations about the consequences with a probability model, and if I choose option 1 I can represent my expectations about the consequences with a different probability model. I can compare the two consequences, then decide which option seems to be better. To make this comparison, I have used a function from elements of A to probability models. A function that takes elements of some set as inputs (which may or

may not be decisions) and returns probability models is a *probability gap model*, and the set of inputs it accepts is a *probability gap*.

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

Higher order probability gap models can be represented by conditional probabilities $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ or pairs of conditional probabilities $\{\mathbb{P}^{\mathsf{X}|\mathsf{W}}, \mathbb{P}^{\mathsf{Z}|\mathsf{WXY}}\}$, which we call *order 1* and *order 2* models respectively. Decision functions in data-driven decision problems correspond to probability gaps in order 2 models, as we discuss in Section 3, which makes this type of model particularly interesting for our purposes. We also require these to be valid, and we define conditions for validity and prove that they are sufficient to ensure that models represented by conditional probabilities can in fact be mapped to base measures on the fundamental probability set.

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

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

this.

2.2 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)**.**

Lemma 2.9. *lem:pushf_kprodGivenaprobabilityspace$(\mu, \Omega, \mathcal{F})$ 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

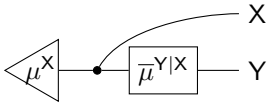
Instead of having probability distributions and Markov kernels as two different kinds of thing, we can identify probability distributions with Markov kernels whose domain is a one element set $\{*\}$.

Definition 2.10 (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 2.11 (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

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

$$\iff \tag{6}$$

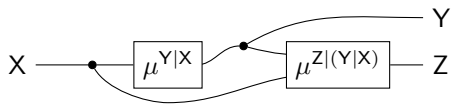
$\mu^{\mathsf{XY}} =$ 

(7)

We define higher order conditionals as "conditionals of conditionals"

Definition 2.12 (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

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 *valid-*

ity such that a collection of valid conditional probabilities recursively combined using $\odot$ is guaranteed to corresponds to a non-empty probability set.

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 2.13 (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 2.11 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 2.19) with respect to probability sets.

Definition 2.14 (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 2.15 (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 $X \nrightarrow Y$ such that $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\{\}}$ is an $X \nrightarrow Y$ disintegration of $\mathbb{P}^{\mathsf{XY}}_{\alpha}$ for all $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$. If no such Markov kernel exists, $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\{\}}$ is undefined.

Definition 2.16 (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 $X \times Y \nrightarrow Z$ 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 $\mathbb{P}^{\mathsf{Z}|\mathsf{YX}}_{\{\}}$.

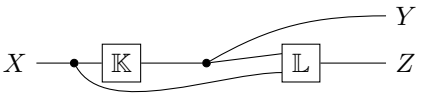
2.6 Semidirect product and almost sure equality

The operation used in Equation 94 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 2.17 (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

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

$$= \tag{11}$$



$$\iff \tag{12}$$

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

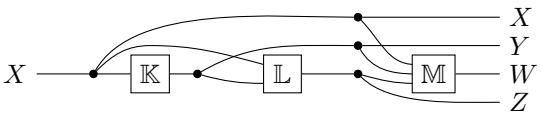
Lemma 2.18 (Semidirect product is associative)**.** *Given* $\mathbb{K} : X \nrightarrow Y$, $\mathbb{L} : Y \times X \nrightarrow Z$ *and* $\mathbb{M} : Z \times Y \times X \nrightarrow W$

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

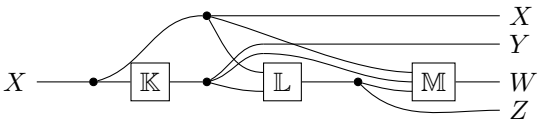
$$\tag{15}$$

Proof.

$$(\mathbb{K} \odot \mathbb{L}) \odot \mathbb{M} = \tag{16}$$



$$= \tag{17}$$



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

□

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 $\mathbb{P}^{\mathsf{X}}_{\alpha}$ with each Markov kernel is identical.

Definition 2.19 (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 2.20 (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 $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$

$$\begin{aligned} \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{K} &= \mathbb{P}_\alpha^{\mathsf{XY}} && (20)\\ &= \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{L} && (21) \end{aligned}$$

□

Lemma 2.21 (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 $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$

$$\begin{aligned} \mathbb{P}_\alpha^{\mathsf{XY}} \odot \mathbb{K} &= (\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}) \odot \mathbb{K} && (23)\\ &= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{K}) && (24)\\ &= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{L}) && (25) \end{aligned}$$

□

Lemma 2.22 (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*

$$\begin{aligned} \mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}} &= \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{XY}} && (26)\\ & && (27) \end{aligned}$$

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

$$\begin{aligned} \mathbb{P}_\alpha^{\mathsf{XYZ}} &= \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_\alpha^{\mathsf{YZ}|\mathsf{X}} && (28)\\ &= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_\alpha^{\mathsf{Z}|\mathsf{YX}}) && (29)\\ &= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}) && (30) \end{aligned}$$

□

2.7 Probability sets defined by marginal and conditional probabilities

So far we have defined probability sets and conditional probabilities as Markov kernels that can sometimes be derived from a probability set. Actually, we are interested in working in the opposite direction: starting with conditional probabilities and working with probability sets defined by them. We need to be a little bit careful in doing this: 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 $\mathsf{Y} : \Omega \to Y$. The reason for this is that some collections of variables cannot have arbitrary conditional probabilities.

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. A *valid distribution* is a distribution associated with a particular variable that defines a nonempty set of base measures on Ω (Theorem 5.12), and *valid conditionals* are a set of conditional probabilities closed under $\odot$ and reducing to valid distributions when conditioning on a trivial variable (Lemma 5.15).

Definition 2.23 (Valid distribution)**.** A valid X probability distribution $\mathbb{P}^{\mathsf{X}}$ is any probability mesure on $\Delta(X)$ such that $\mathsf{X}^{-1}(A) = \emptyset \implies \mathbb{P}^{\mathsf{X}}(A) = 0$ for all $A \in \mathcal{X}$.

Definition 2.24 (Valid conditional)**.** Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ a *valid* $\mathsf{Y}|\mathsf{X}$ *conditional probability* $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ is a Markov kernel $X \nrightarrow Y$ such that it assigns probability 0 to contradictions:

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

Definition 2.25 (Probability set defined by a valid conditional)**.** If $\mathbb{P}_{\{\}}$ is a probability set such that there is a valid conditional probability $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} : X \nrightarrow Y$ and for every $\mu \in \Delta(\Omega)$ such that $\mu^{\mathsf{Y}|\mathsf{X}} \overset{\mu}{\in} \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$, we say $\mathbb{P}_{\{\}}^{\overline{\mathsf{Y}|\mathsf{X}}} := \mathbb{P}_{\{\}}$ is the probability set defined by $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$.

Suppose we have some collection of Markov kernels that we want to interpret as conditional probabilities $\{\mathbb{P}_i^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in [n]\}$ which we want to define a probability set by recursively taking the semidirect product $\mathbb{P}_1^{\mathsf{X}_1} \odot (\mathbb{P}_2^{\mathsf{X}_2|\mathsf{X}_1} \odot ...)$. It is sufficient that each $\mathbb{P}_i^{\mathsf{X}_i|\mathsf{X}_{[i-1]}}$ is valid for the resulting probability set to be nonempty (Lemma 5.15).

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 $\mathbb{K} : X \twoheadrightarrow Y$, $\mathbb{K}$ 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 5.17).

2.8 Probability gap models

For reasoning about decisions, we don't just want a set of models that could explain what is going on. What we want is a function that maps choices to "outcome" probability models. In many cases, there might be features that all of the outcome models share – for example, there might be some variables that are not affected by any choice, and so their marginal distribution is the same in every outcome model. There are also some other features that are entirely determined by the choice we make. For example, if there is a variable $\mathsf{D} : \Omega \to D$ representing the choice we make, then if we choose option $d \in D$ we must have $\mathbb{P}^{\mathsf{D}}(\{d\}) = 1$.

Here, by "property", we mean marginal or conditional probabilities. We suppose that we are able to construct models such that "properties common to all outcome models" and "properties determined by choices" together represent everything we can say about the appropriate model for our decision problem. That is, we can define a probability set corresponding to all outcome models and, for each choice, a probability set corresponding to all models consistent with the things known to be fixed by that choice, and then the result of picking a certain choice is the intersection of the probability set for all outcome models and the probability set associated with that choice.

We call this kind of map a *probability gap model* (the terminology is from Hájek (2019), though our meaning is a little different). The set of all outcome models represents most of our knowledge relevant to the outcome, but there's a gap – it doesn't say which choice we will eventually make. The set of choices is the collection of different ways that the gap could be filled.

We don't have an axiomatic justification for choosing this representation. There are two considerations motivating it: first, it allows us to recover standard representations of decision problems and standard kinds of causal models, and secondly the formulation in terms of probability sets means that probability gap models are in many ways similar to ordinary probability models.

- A fixed probability set $\mathbb{P}_{\{\}} \subset \Delta(\Omega)$ which we call the *model*
- A collection of probability sets $A \subset \Delta(\Omega)$ that we call *choices*
- A map $\mathbb{P}_{\square} : A \to \mathcal{P}(\Delta(\Omega))$ defined by $\mathbb{P}_{\alpha} := \mathbb{P}_{\square}(\alpha) = \mathbb{P}_{\{\}} \cap \alpha$

We require that the choices are compatible with the model in the sense that $\mathbb{P}_{\{\}} \cap \alpha \neq \emptyset$ for all $\alpha \in A$. Here, we will limit our attention to a a particular type

of probability gap model, where we define the probability set $\mathbb{P}_{\{\}}$ is defined by a conditional probability and each choice is defined by a marginal probability relative to the same variable.

Definition 2.26 (Conditional probability model)**.** A *conditional probability model* $\mathbb{P}_{\square}$ is a probability gap model $(\mathbb{P}_{\{\}}^{\overline{\mathsf{Y}|\mathsf{X}}}, A)$ such that each $\alpha \in A$ is some probability set defined by an X-valid marginal probability $\alpha^{\overline{\mathsf{X}}}$.

We will compute the intersection $\mathbb{P}_\alpha$ between the model $\mathbb{P}_{\{\}}$ and a choice $\alpha \in A$ as the probability set $\mathbb{P}_\alpha^{\overline{\mathsf{XY}}}$ such that:

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

This is justified by Lemma 5.13, which says that the probability set defined by Equation 32 is equivalent to the intersection of α and $\mathbb{P}_{\{\}}$.

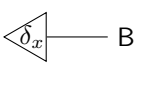
If the conditional probability $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ and all the marginal probabilities $\alpha^{\overline{\mathsf{X}}}$ are valid, then by Lemma 5.15 $\mathbb{P}_{\{\}} \cap \alpha \neq \emptyset$ for all $\alpha \in A$. Thus validity of all the individual parts is enough to ensure 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.

2.9 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

$$\nu^{\mathsf{B}|\mathsf{WH}} = \tag{33}$$



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

$$(\mathsf{B}, \mathsf{W}, \mathsf{H}) \bowtie \{(x, w, h)\} = \{\omega | (\mathsf{W}, \mathsf{H})(\omega) = (w, h), \mathsf{B}(\omega) = \frac{w}{h^2}\} \tag{34}$$

$$= \emptyset \tag{35}$$

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.

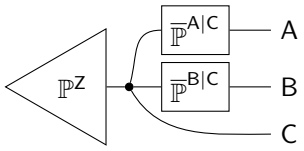
Validity rules out conditional probabilities like 33. We guess that in many cases this condition may either be trivial or unconsiously taken into account when constructing conditional probabilities. However, if we are not cognizant of the conditional our model depends on, we may inadvertently propose a model that depends on invalid conditional probabilities. For example, the conditional probability 33 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.

2.9.1 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 5.2.

Definition 2.27 (Conditional independence with respect to a probability set)**.** For a *probability set* $\mathbb{P}_{\{\}}$ and variables $\mathsf{A}, \mathsf{B}, \mathsf{Z}$, we say B is conditionally independent of A given C, written $\mathsf{B} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{A}|\mathsf{C}$, if

$$\mathbb{P}^{\mathsf{ABC}}_{\{\}} = \tag{36}$$

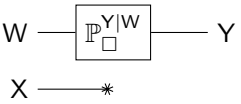


Cho and Jacobs (2019) have shown that this definition coincides with the standard notion of conditional independence for a particular probability model. In particular, it satisfies the *semi-graphoid axioms*.

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

Theorem 2.28. *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*

$$\mathbb{P}^{\mathsf{Y}|\mathsf{WX}} = \tag{37}$$



$$\iff \mathbb{P}^{\mathsf{Y}|\mathsf{WX}}(y|w,x) = \mathbb{P}^{\mathsf{Y}|\mathsf{W}}(y|w) \tag{38}$$

Proof. See Cho and Jacobs (2019). □

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

2.10 Curried Markov kernels

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

$$\lambda\mathbb{K}((y_i)_{i\in X}) = \prod_{i\in X} \mathbb{K}(y_i|i) \tag{39}$$

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

$$\mathbb{K} = (\lambda\mathbb{K} \otimes \mathrm{id}_X)\mathbb{F}_{\mathrm{ev}} \tag{40}$$

Unlike the case of function currying, $\lambda\mathbb{K}$ is not the unique Markov kernel for which 40 holds. In fact, we can substitute any $\mathbb{L}$ such that, for any $i \in X$

$$\sum_{y_{\{i\}^C \in Y^{|X|-1}}} \mathbb{L}((y_i)_{i\in X}) = \mathbb{K}(y_i|i) \tag{41}$$

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

$$\mathsf{Y} \overset{a.s.}{=} \mathrm{ev}(\mathsf{Y}^X, \mathsf{X}) \tag{42}$$

Then

$$(\mathbb{P}^{\mathsf{Y}^X} \otimes \mathrm{id}_X)\mathbb{F}_{\mathrm{ev}} \tag{43}$$

is some Markov kernel $\mathbb{K} : X \nrightarrow Y$, which is equal to $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ if $\mathsf{Y}^X \perp\!\!\!\perp |\mathsf{X}$. However, potential outcomes models typically do not explain what the kernel $\mathbb{K}$ represents, and instead offer a definition of the variable Y^X. For $x \in X$, the component Y^x of Y^X is usually said to express "the outcomes that would have been observed, if X was x".

Our original motivating question was "when are potential outcomes well-defined?". We're not actually going to try to answer this question, because our

aim is not to tell people using potential outcomes how to do it. Furthermore, that question invites controversy we are not particularly interested in joining; Dawid (2000) and Richardson and Robins (2013) have both argued that it is better to use equivalence classes of potential outcomes models induced by a criterion of distinguishability by experiment, while Pearl (2009) advocates for models that can make finer distinctions than this.

However, given a probability gap model $\mathbb{P}_\square$, we do have a natural notion of the well-definedness of a conditional probability $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{X}}$ – it is well-defined when $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}$ is equal for all α (Definition 2.16). Furthermore, the formal conditions that guarantee the existence of such a conditional probability very closely resemble the *stable unit treatment value assumption* (SUTVA), which is said to be necessary for the existence of potential outcomes 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.

The added emphasis is ours. In the next section, we offer formal criteria that correspond to these two statements.

start again here

3 Decision theoretic causal inference

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

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

We do not have a comparable axiomatisation of preferences that yield a representation of decision problems in terms of utility and gappy probability. Such an undertaking could potentially clarify some choices that can be made in

setting up a gappy probability model of decision making, but it is the subject of future work. Instead, we suppose that we are satisfied with a particular probabilistic model of a decision problem, based on convention rather than axiomatisation.

3.1 Decision problems

Suppose we have an observation process $\mathcal{X}$, modelled by X taking values in X (we are *informed*). Given an observation $x \in X$, we suppose that we can choose a decision from a known set D (the set of decisions is *transparent*), and we suppose that choosing a decision results in some action being taken in the real world. As with processes of observation, we will mostly ignore the details of what "taking an action" involves. The process of choosing a decision that yields an element of D is a decision making process $\mathcal{D}$ modelled by D. We might be able to introduce randomness to the choice, in which case the relation between X and D may be stochastic. We will assume that there is some $\mathcal{Y}$ modelled by Y such that $(\mathsf{X}, \mathsf{D}, \mathsf{Y})$ tell us everything we want to know for the purposes of deciding which outcomes are better than others.

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

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

We consider an additional kind of gap in our probability model. The nature of this gap is: we don't know exactly which order 2 model $(\mathbb{P}_\square^{\mathsf{X},\mathbb{P}_\square^{\mathsf{Y}}|\mathsf{XD}}, A)$ we "ought" to use. To represent this gap we include an unobserved variable H, the *hypothesis*. We can interpret H as expressing the fac that, if we knew the value of H then we would know that our decision problem was represented by a unique order 2 model $(\mathbb{P}_{h,\square}^{\mathsf{X},\mathbb{P}_{h,\square}^{\mathsf{Y}}|\mathsf{XD}}, A)$. However, H is not known and in fact we do not know how to determine H (this is the nature of an *unoserved* variable – there is

no process available to find the value it yields). Our model is thus given by

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

.

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

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

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

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

I need to update the proof for this claim

3.2 Decisions as measurement procedures

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

Suppose that we are only contemplating decision functions that map deterministically to D. Suppose furthermore that we will D according to a model $\mathbb{P}_{\square}$, a utility function on $X \times D \times Y \to \mathbb{R}$ and a decision rule which is a function f from models, utility functions and decision rules to decisions. Note that models, utility functions and decision rules are all well-defined mathematical objects. If we are confident that our choice will in the end be an element of a well-defined set of objects of the appropriate type, then we are positing that we have a "measurement procedure" $\mathcal{M}$ that yield models, utilities and decision rules. If so, $f \circ \mathcal{M}$ – that is, the function that yields a decision – is itself a measurement procedure. This is what is unique about decisions: proposing a complete decision problem with models, utilities and decision rules, defines a

measurement procedure for decisions. Other quantities of interest do not seem to have this property – we *require* a measurement process for observations in order to make the whole setup work, but we do not *define* it in the course of setting up a model for our decision problem.

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

3.3 Causal models similar to see-do models

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

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

> A fundamental feature of the DT approach is its consideration of the relationships between the various probability distributions that govern different regimes of interest. As a very simple example, suppose that we have a binary treatment variable T, and a response variable Y. We consider three different regimes [...] the first two regimes may be described as interventional, and the last as observational.

The difference between the model described here and a see-do model is that a see-do model uses different variables X and Y to represent observations and consequences, while Dawid's model uses the same variable (T, Y) to represent outcomes in interventional and observational regimes. In this work we associate one observed variable with each measurement process, while in Dawid's approach (T, Y) seem to be doing double duty, representing mesurement processes carried out during observations and after taking action. This can be thought of as the causal analogue of the difference betwen saying we have a sequence $(\mathsf{X}_1, \mathsf{X}_2, \mathsf{X}_3)$ of observations independent and identically distributed according to $\mu \in \Delta(X)$ and saying that we have some observations distributed according to $\mathbb{P}^{\mathsf{X}} \in \Delta(X)$. People usually understand what is meant by the latter, but if

one is trying to be careful the former is a more precise statement of the model in question.

Heckerman and Shachter (1995) also explore a decision theoretic approach to causal inference. Our approach is quite close to their approach if we identify what we call hypotheses with what they call states and allow for probabilistic dependence between states, decisions and consequences. It is an open question whether their notion of limited unresponsiveness corresponds to any notion of conditional independence in our work.

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

3.4 See-do models and classical statistics

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

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

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

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

Suppose we have a see-do model $\mathbb{P}^{\mathsf{X}|\mathsf{H}\square\mathsf{Y}|\mathsf{D}}$ with $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}} \mathsf{X}|(\mathsf{H},\mathsf{D})$, and suppose that the random variable Y is a "negative utility" function taking values in $\mathbb{R}$

for which *low* values are considered desirable. Define a loss $l : H \times D \to \mathbb{R}$ by $l(h,d) = \sum_{y \in \mathbb{R}} y \mathbb{P}^{\mathsf{Y}|\mathsf{HD}}(y|h,d)$, we have

$$\mathbb{E}_{\mathbb{P}_\alpha}[\mathsf{Y}|h] = \sum_{x \in X} \sum_{d \in D} \sum_{y \in Y} \mathbb{P}^{\mathsf{X}|\mathsf{H}}(x|h) \mathbb{Q}_\alpha^{\mathsf{D}|\mathsf{X}}(d|x) \mathbb{P}^{\mathsf{Y}|\mathsf{HD}}(y|h,d) \tag{44}$$

$$= \sum_{x \in X} \sum_{d \in D} \mathbb{P}^{\mathsf{X}|\mathsf{H}}(x|h) \mathbb{Q}_\alpha^{\mathsf{D}|\mathsf{X}}(d|x) l(h,d) \tag{45}$$

$$= R(h,\alpha) \tag{46}$$

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

4 Syntax and semantics of causal consequences

Causal Bayesian networks and potential outcomes employ different naming conventions to distinguish "causal effects" from "simple correlations". Causal Bayesian networks write $P(\mathsf{Y}|do(\mathsf{X}))$ and $P(\mathsf{Y}|\mathsf{X})$, while potential outcomes distinguishes $P(\mathsf{Y}|\mathsf{X})$ from $x \mapsto P(\mathsf{Y}^x)$. If we are not going to worry too much about details of interpretation, we can interpret the expression $P(\mathsf{Y}|\mathsf{X})$ as expressing something like this: there is an objective probability $P(\mathsf{Y},\mathsf{X})$ that describes a sequence of independent and identically distributed observations, and $P(\mathsf{Y}|\mathsf{X})$ is a disintegration of this probability. The existence of an objective probability $P(\mathsf{Y},\mathsf{X})$ can be justified by an assumption that the sequence of observations should be modeled exchangeably.

We pursue a similar line of thinking with respect to understanding causal consequences like $P(\mathsf{Y}|do(\mathsf{X}))$ or $x \mapsto P(\mathsf{Y}^x)$. We assume that "causal consequences" are conditional probabilities of the form $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{DH}}$ where Y is an outcome, D is some decision, H is a hypothesis and $\mathbb{P}_\square$ is a probability gap model. Our interest is in understanding what causal consequences are *from the point of view of someone choosing a decision function*. We do not address the question of how they may be inferred from observed data.

We show that conditional probability models that are *causally contractible* with respect to a sequence of decisions and a corresponding sequence of outcomes are representible by mixtures of "objective but unknown" conditional probabilities. This is analogous to De Finetti's theorem that shows exchangeable probability distributions are representable by mixtures of "objective but unknown" independent and identically distributed probability distributions. A similar argument to ours is found in Dawid (2020).

We also consider the question of when causal contractibility could be supposed to hold. This is a subtle question, as the answer appears to differ for situations that are quite similar. For example, consider:

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 considers the available research and formulates a general sense of whether or not she'll treat each one, which she quantifies with $\mathbb{P}_\alpha^{\mathsf{D}_1\mathsf{D}_2}$

2. As before, but prior to seeing the patients she considers the available research and decides to treat on the basis of applying a function to a random number generator with known characteristics. The choice of function and random number generator allow her to quantity probability of treatment with $\mathbb{P}_\alpha^{\mathsf{D}_1\mathsf{D}_2}$

I removed the discussion of probability combs for simplicity, so I have not considered policies for treatment that depend on earlier experiments in the examples above

We will argue that Alice could reasonably assume causal contractibility in the second case but not the first. While we are unable to offer a general theory of when causal contractibility holds, we note that an apparently key difference between the two situations is that in the first case the "decision" D_1 is indeterministic for some α, though D_2 is deterministic, while in the second case both D_1 and D_2 are determinstic functions.

4.1 Repeatable experiments

A conditional probability model $(\mathbb{P}_\square^{\overline{\mathsf{Y}|\mathsf{D}}}, A)$ is a model of a sequential experiment if $\mathsf{Y} := \mathsf{Y}_M = (\mathsf{Y}_i)_{i\in M}$ and $\mathsf{D} := \mathsf{D}_M = (\mathsf{D}_i)_{i\in M}$ for some index set M. We say that Y_i is the consequence corresponding to the decision D_i for all $i \in M$. We identify a "causal consequence" with a conditional probability of the form $\mathbb{P}_\square^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i}$, where H is a hypothesis that is deterministically identical for every i. Causal consequences do not generally exist, see Definition 2.16.

If $(\mathbb{P}_\square^{\overline{\mathsf{Y}|\mathsf{D}}}, A)$ represents a sequential experiment, we might guess that causal consequences exist if the experiment is in some sense "repeatable". We consider two precise notions of repeatability. The first condition is *commutativity of exchange*, which is the assumption that swapping the choices that we apply at each step and then applying the corresponding inverse swap to consequences leaves the model unchanged. The second condition is *commutativity of marginalisation* – if we perform the whole experiment multiple times, making the same choice D_i at any point i gets the same results, regardless of what other choices are made.

Commutativity of exchange is similar to the condition of *post-treatment exchangeability* found in Dawid (2020), and commutativity of marginalisation is similar to the stable unit treatment distribution assumption (SUTDA) in the same, as well as the "no interference" part of the stable unit treatment value assumption (SUTVA) with which it shares a name. Commutativity of exchange is also very similar to the exchangeability assumption of GREENLAND and ROBINS (1986) for further discussions of exchangeability in the context of causal modelling, and note that both authors consider exchanging to be an

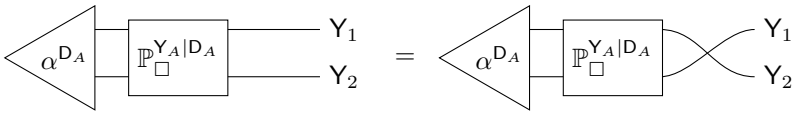
operation that alters which person receives which treatment. The assumption of exchangeability found in Banerjee et al. (2017) can also be regarded as similar to commutativity of exchange.

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

Commutativity of exchange 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):



(47)

Commutativity of exchange (swapping choices $\sim$ swapping labels):

commutativity of exchange (48)

Commutativity of exchange is a property of probability gap models, not a property of fixed probability model for which there is no analogue of "attaching a different choice" in that case.

—-end not sure where to put——

More precisely, a conditional probability model "commutes with exchange" if applying any finite permutation to blind choices or separately applying the corresponding permuation to consequences each yields the same result. We can apply the exchange "before" multiplying by the conditional $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{D}}$ or after it and we get the same result.

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 (Commutativity of exchange)**.** Suppose we have a sample space $(\Omega, \mathcal{F})$ and a conditional probability model $(\mathbb{P}_\square^{\overline{\mathsf{Y}|\mathsf{D}}}, A)$ with $\mathsf{Y} = \mathsf{Y}_M$, $\mathsf{D} = \mathsf{D}_M$, $M \subseteq \mathbb{N}$. If, for any two decision rules $\alpha^{\overline{\mathsf{D}}}, \beta^{\overline{\mathsf{D}}} \in A$,

$$\alpha^{\mathsf{D}} \odot \mathrm{swap}_{\rho(\mathsf{D})}\mathbb{P}_\square^{\mathsf{Y}|\mathsf{D}} = \alpha^{\mathsf{D}} \odot \mathbb{P}_\square^{\mathsf{Y}|\mathsf{D}}\mathrm{swap}_{\rho(\mathsf{Y})} \tag{49}$$

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

A do model is non interfering if it gives identical results for identical subsequences of different choices when we limit our attention to the corresponding subsequences of consequences. For example, if we have $\mathsf{D} = (\mathsf{D}_1, \mathsf{D}_2, \mathsf{D}_3)$ and $\mathsf{Y} = (\mathsf{Y}_1, \mathsf{Y}_2, \mathsf{Y}_3)$ and $\alpha^{\mathsf{D}_1\mathsf{D}_3} = \mathbb{P}_\beta^{\mathsf{D}_1\mathsf{D}_3}$ 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 (Commutativity of marginalisation)**.** Suppose we have a sample space $(\Omega, \mathcal{F})$ and a conditional probability model $(\mathbb{P}_\square^{\overline{\mathsf{Y}|\mathsf{D}}}, A)$ with $\mathsf{Y} = \mathsf{Y}_M$, $\mathsf{D} = \mathsf{D}_M$, $M \subseteq \mathbb{N}$. For any $S = (s_i)_{i\in Q}$, $Q \subset M$, and $i < j \implies p_i < p_j \ \& \ q_i < q_j$, let $\mathsf{D}_S := (\mathsf{D}_i)_{i\in S}$ and $\mathsf{D}_T := (\mathsf{D}_i)_{i\in T}$. If for any $\alpha, \beta \in R$

$$\mathbb{P}_\alpha^{\mathsf{D}_S} = \mathbb{P}_\beta^{\mathsf{D}_S} \tag{50}$$

$$\implies \mathbb{P}_\alpha^{(\mathsf{D}_i,\mathsf{Y}_i)_{i\in S}} = \mathbb{P}_\beta^{(\mathsf{D}_i,\mathsf{Y}_i)_{i\in S}} \tag{51}$$

then $\mathbb{P}_\square$ *commutes with marginalisation.*

Neither condition implies the other.

Lemma 4.4. *Commutativity of exchange does not imply commutativity or vise versa.*

Proof. Suppose $D = Y = \{0, 1\}$ and we have a conditional probability model $(\mathbb{P}_\square^{\overline{\mathsf{Y}|\mathsf{D}}}, A)$ 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

$$\mathbb{P}_\square^{\mathsf{Y}_1\mathsf{Y}_2|\mathsf{D}_1\mathsf{D}_2}(y_1, y_2|d_1, d_2) = [\![(y_1, y_2) = (d_1 + d_2, d_1 + d_2)]\!] \tag{52}$$

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 $\mathbb{P}_\square$ does not commute with marginalisation. However, taking $(d_i, d_j) := \delta_{d_i d_j} \in A$,

$$\mathbb{P}_{d_2,d_1}^{\mathsf{Y}_1\mathsf{D}_1\mathsf{Y}_2\mathsf{D}_2}(y_1, d_1, y_2, d_2) = [\![(y_1, y_2) = (d_2 + d_1, d_2 + d_1)]\!] \tag{53}$$

$$= [\![(y_2, y_1) = (d_1 + d_2, d_1 + d_2)]\!] \tag{54}$$

$$= \mathbb{P}_{d_1,d_2}^{\mathsf{Y}_1\mathsf{D}_1\mathsf{Y}_2\mathsf{D}_2}(y_2, d_2, y_1, d_1) \tag{55}$$

so $\mathbb{P}_\square$ commutes with exchange.
Alternatively, suppose the same setup, but define $\mathbb{P}_\square$ instead by, for all $\alpha \in A$

$$\mathbb{P}_\square \mathsf{Y}_1\mathsf{Y}_2|\mathsf{D}_1\mathsf{D}_2(y_1, y_2|d_1, d_2) = [\![(y_1, y_2) = (0, 1)]\!] \tag{56}$$

Then $\mathbb{P}_\square$ commutes with marginalisation. If $\mathbb{P}_\alpha^{\mathsf{D}_S} = \mathbb{P}_\beta^{\mathsf{D}_S}$ for $S \subset \{0, 1\}$ then

$$\mathbb{P}_\alpha^{\mathsf{Y}_S\mathsf{D}_S}(y_s, d_s) = \sum_{y_2' \in \{0,1\}^{S^C}} [\![(y_1, y_2) = (0, 1)]\!] \mathbb{P}_\alpha^{\mathsf{D}_S}(d_s) \tag{57}$$

$$= \mathbb{P}_\beta^{\mathsf{Y}_S\mathsf{D}_S}(y_s, d_s) \tag{58}$$

but not exchange. For all $\alpha, \beta \in A$:

$$\mathbb{P}_\alpha \mathsf{Y}_1 \mathsf{Y}_2(y_1, y_2) = [\![(y_1, y_2) = (0, 1)]\!] \tag{59}$$
$$\neq \mathbb{P}_\beta \mathsf{Y}_1 \mathsf{Y}_2(y_2, y_1) \tag{60}$$

□

Although commutativity of marginalisation seems to be a bit like non-interference – the marginal distribution I get for Y_i depends only on the decision D_i – it still allows for some models in which we seem to have interference of a kind. For example: in the first experiment I flip a coin and decide either to pass the results to the second experiment ($\mathsf{D}_1 = 0$) or flip another coin and pass those results second experiment ($\mathsf{D}_1 = 1$). In the second I either copy the results I have been given ($\mathsf{D}_2 = 0$) or invert them ($\mathsf{D}_2 = 1$). Then

- The marginal distribution of both experiments is Bernoulli(0.5) no matter what choices I make, so it satisfies Definition 4.3
- Nevertheless, the choice for the first experiment seems to "affect" the result of the second experiment (affect in quotes because it is an intuitive judgement, not a formal property)

Here we are most interested in the conjunction of these assumptions, a condition we call *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.

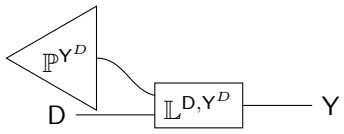
4.2 Causal consequences exist if the model is causally contractible

The main result in this section is Theorem 4.7 which shows that a conditional probability model $\mathbb{P}_\square$ is causally contractible if and only if it can be represented as the product of a distribution over hypotheses $\mathbb{P}_\square^{\mathsf{H}}$ and a collection of identical conditional probabilities $\mathbb{P}_\square^{\mathsf{Y}_1|\mathsf{D}_1\mathsf{H}}$. This can be interpreted as expressing the idea that all $(\mathsf{Y}_i, \mathsf{D}_i)$ pairs share a canonical but unknown "consequence function" $D \twoheadrightarrow Y$. As discussed in Section 2.10, the existence of such a consequence function implies the existence of a common unknown curried consequence function. Curried consequence functions look very similar to potential outcomes models, but they don't necessarily support any counterfactual interpretation.

Lemma 4.6 (Exchangeable curried representation)**.** *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 con-*

tractible if and only if

$$\mathbb{P}^{\mathsf{Y}|\mathsf{D}}_{\square} = \tag{61}$$



$$\iff \tag{62}$$

$$\mathbb{P}^{\mathsf{Y}|\mathsf{D}}_{\square}(y|d) = \mathbb{P}^{(\mathsf{Y}^D_{d_i i})_{\mathbb{N}}}(y) \tag{63}$$

Where $\mathbb{P}^{\mathsf{Y}^D}$ *is an exchangeable probability measure on* $Y^{D\times\mathbb{N}}$*, for convenience we extend the sample space with the random variable* $\mathsf{Y}^D := (\mathsf{Y}^D_{ij})_{i\in D, j\in\mathbb{N}}$ *and* $\mathbb{L}^{\mathsf{D},\mathsf{Y}^D}$ *is the Markov kernel associated with the lookup function*

$$l : D^{\mathbb{N}} \times Y^{D\times\mathbb{N}} \to Y \tag{64}$$

$$((d_i)_{\mathbb{N}}, (y_{ij})_{i\in D, j\in\mathbb{N}}) \mapsto y_{d_i i} \tag{65}$$

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

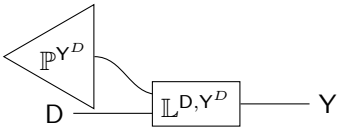
Define

$$\mathbb{P}^{\mathsf{Y}^D}((y_{ij})_{D\times\mathbb{N}}) := \mathbb{P}^{\mathsf{Y}}_e((y_{|D|i+j})_{i\in D, j\in\mathbb{N}}) \tag{66}$$

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

$$\mathbb{Q} : D \nrightarrow Y \tag{67}$$

$$\mathbb{Q} := \tag{68}$$



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

$$\sum_{y\in\mathsf{Y}^{-1}(y_A)} \mathbb{Q}(y|d) = \sum_{y\in\mathsf{Y}^{-1}(y_A)} \mathbb{P}^{(\mathsf{Y}^D_{d_i i})_A}(y) \tag{69}$$

$$= \sum_{y\in\mathsf{Y}^{-1}(y_A)} \mathbb{P}^{(\mathsf{Y}_{|D|d_i+i})_A}_e(y) \tag{70}$$

$$= \mathbb{P}^{\mathsf{Y}_B}_e(y_A) \tag{71}$$

$$= \mathbb{P}^{\mathsf{Y}_A}_d(y_A) \qquad \text{by causal contractibility} \tag{72}$$

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

$$\mathbb{Q}(y|d) = \mathbb{P}_d^{\mathsf{Y}}(y) \tag{73}$$

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

$$\mathbb{Q}(y|d) = \mathbb{P}_d^{\mathsf{Y}|\mathsf{D}}(y|d) \tag{74}$$

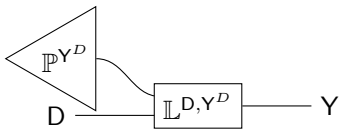
And because $\mathbb{P}_d^{\mathsf{Y}|\mathsf{D}}(y|d)$ is unique for all $d \in D^{\mathbb{N}}$ and $\mathbb{P}^{\mathsf{Y}|\mathsf{D}}$ exists by assumption

$$\mathbb{P}^{\mathsf{Y}|\mathsf{D}} = \mathbb{Q} \tag{75}$$

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

$$\begin{aligned}
\mathbb{P}^{\mathsf{Y}_S^D} &= \mathbb{P}_e^{\mathsf{Y}_{\rho(S)}}) && (76)\\
&= \mathbb{P}_e^{\mathsf{Y}_{\rho(T)}}) \quad \text{by contractibility of } \mathbb{P} \text{ and the equality } e_{\rho(S)} = e_{\rho(T)} && (77)\\
&= \mathbb{P}^{\mathsf{Y}_T^D} && (78)
\end{aligned}$$

If: Suppose

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

(79)

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,

$$\begin{aligned}
\mathbb{P}^{\mathsf{Y}_S|\mathsf{D}}(y_S|d) &= \sum_{y_S^D \in Y^{|D|\times|S|}} \mathbb{P}^{\mathsf{Y}_S^D}(y_S^D)\mathbb{L}^{\mathsf{D}_S,\mathsf{Y}^S}(y_S|d, y_S^D) && (80)\\
&= \sum_{y_S^D \in Y^{|D|\times|T|}} \mathbb{P}^{\mathsf{Y}_T^D}(y_S^D)\mathbb{L}^{\mathsf{D}_S,\mathsf{Y}^S}(y_S|d, y_S^D) \quad \text{by contractibility of } \mathbb{P}^{\mathsf{Y}_T^D} && (81)\\
&= \mathbb{P}^{\mathsf{Y}_T|\mathsf{D}}(y_S|d) && (82)
\end{aligned}$$

□

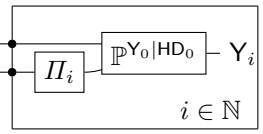
The curried representation of Lemma 4.6 does not need to support an interpretation as a distribution of potential outcomes. For example, consider a series of bets on fair coinflips – in this case, the consequence Y_i is uniform on $\{0,1\}$ for any decision D_i. Tha $D = Y = \{0,1\}$ and $\mathbb{P}_\alpha^{\mathsf{Y}_n}(y) = \prod_{i\in[n]} 0.5$ for all n, $y \in Y^n$, $\alpha \in R$. Then the construction in Lemma 4.6 yields $\mathbb{P}^{\mathsf{Y}_i^D}(y_i^D) = \prod_{j\in D} 0.5$ for all $y_i^D \in Y^D$. That is, Y_i^0 and Y_i^1 are independent and uniformly distributed. However, if we wanted Y_i^0 to represent "what would happen if I bet on outcome 0 on turn i" and Y^1 to represent "what would happen if I bet on outcome 1 on turn i", then it seems that we ought to have $\mathsf{Y}_i^0 = 1 - \mathsf{Y}_i^1$.

We could suppose that Lemma 4.6 provides necessary but not sufficient conditions for the existence of a potential outcomes representation of a conditional probability model. However, it doesn't seem to succeed at that either. We note, for example, that Rubin (2005) does not assume that the distribution of potential outcomes is exchangeable. A non-exchangeable $\mathbb{P}^{\mathsf{Y}^D}$ does not induce a causally contractible conditional probability model, and at the same time commutativity with marginalisation is not sufficient for a conditional probability model to support a curried representation in the sense of Lemma 4.6. What seems to be missing is an additional assumption that consequences are mutually independent of one another given the associated decision.

We can also represent contractible conditional probability models repeated copies of an unknown "consequence function", a Markov kernel that maps from decisions to probability distributions over consequences, coupled by a common hypothesis H.

Theorem 4.7. *Suppose we have a fundamental probability set Ω and a do model $(\mathbb{P}, \mathsf{D}, \mathsf{Y}, R)$ such that $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$. $\mathbb{P}$ is causally contractible if and only if there exists some $\mathsf{H} : \Omega \to H$ such that $\mathbb{P}^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i}$ exists for all $i \in \mathbb{N}$ and*

$$\mathbb{P}^{\mathsf{Y}|\mathsf{H}\mathsf{D}} = \quad \text{(83)}$$



$$\iff \quad \text{(84)}$$

$$\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}} \mathsf{Y}_{\mathbb{N}\setminus i}, \mathsf{D}_{\mathbb{N}\setminus i} | \mathsf{H}\mathsf{D}_i \qquad \forall i \in \mathbb{N} \quad \text{(85)}$$

$$\wedge\mathbb{P}^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i} = \mathbb{P}^{\mathsf{Y}_0|\mathsf{H}\mathsf{D}_0} \qquad \forall i \in \mathbb{N} \quad \text{(86)}$$

Proof. We make use of Lemma 4.6 to show that we can represent the conditional probability as an exchangeable tabular probability distribution. We then use the property of exchangeability of the columns of that distribution in conjunction with De Finetti's theorem to derive the result. □

4.3 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.

4.4 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)
  $\mathcal{Y}_2 \leftarrow$ 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

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
  $(\mathcal{D}_1, \mathcal{D}_2) \leftarrow$ decisions(α)
  $(\mathcal{D}_1^{\text{swap}}, \mathcal{D}_2^{\text{swap}}) \leftarrow (\mathcal{D}_2, \mathcal{D}_1)$
  $\mathcal{Y}_1^{\text{swap}} \leftarrow$ apply($\mathcal{D}_1^{\text{swap}}$, patientA)
  $\mathcal{Y}_2^{\text{swap}} \leftarrow$ apply($\mathcal{D}_2^{\text{swap}}$, 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
  $(\mathcal{D}_1, \mathcal{D}_2) \leftarrow$ decisions(α)
  $\mathcal{Y}_2 \leftarrow$ 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

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$\leftarrow$ inspect(patient A)
  patient B knowledge$\leftarrow$ 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$\leftarrow$ inspect(patient B)
  patient A knowledge$\leftarrow$ 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$\leftarrow$ inspect(patient A)
patient B knowledge$\leftarrow$ 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.

4.5 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 yields an equivalent problem
- 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 the determinism assumption undermines the argument.

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

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

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

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

By Corollary 5.11 and the existence of $\mathbb{P}^{\mathsf{Y}_i\mathsf{X}_i|\mathsf{HD}_i}$ for all $i \in M$, $\mathbb{P}_{\square}^{\mathsf{Y}_i|\mathsf{HX}_i}$ exists for all i. Furthermore, because $\mathbb{P}^{\mathsf{Y}_i\mathsf{X}_i|\mathsf{HD}_i} = \mathbb{P}^{\mathsf{Y}_j\mathsf{X}_j|\mathsf{HD}_j}$ for all $i, j \in M$, $\mathbb{P}_{\square}^{\mathsf{Y}_i|\mathsf{HX}_i} = \mathbb{P}_{\square}^{\mathsf{Y}_j|\mathsf{HX}_j}$ for all $i, j \in M$. □

If the condition $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_\square} \mathsf{D}_i|\mathsf{HX}_i$ 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 $\mathbb{P}_{\square}^{\mathsf{Y}_0|\mathsf{T}_0\mathsf{H}}$.

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.

4.6 Intersubjective causal consequences

While the assumption of causal contractibility itself does not depend on any notion of subjectivity, our discussion of the appliccability 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 **??**, for any $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$

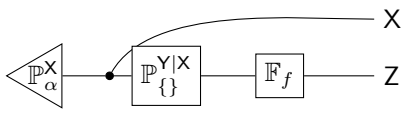
$$\mathbb{P}_\alpha^{\mathsf{XZ}} = \mathbb{P}_\alpha^{\mathsf{XY}} \mathbb{F}_{\mathrm{id}_X \otimes f} \tag{87}$$

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

$$\begin{aligned} \mathbb{F}_{\mathrm{id}_X \otimes f}(A \times B | x, y) &= \delta_x(A) \delta_{f(y)}(B) && (88) \\ &= \mathbb{F}_{\mathrm{id}_X}(A|x) \otimes \mathbb{F}_f(B|y) && (89) \\ \implies \mathbb{F}_{\mathrm{id}_X \otimes f} &= \mathbb{F}_{\mathrm{id}_X} \otimes \mathbb{F}_f && (90) \end{aligned}$$

Thus

$$\mathbb{P}_\alpha^{\mathsf{XZ}} = (\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}) \mathbb{F}_{\mathrm{id}_X} \otimes \mathbb{F}_f \tag{91}$$



$$= \tag{92}$$

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 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

$$n(\mathsf{Y}^{-1}(y)|x, y) = 1 \tag{93}$$

$$m(\mathsf{Y}^{-1}(A) \cap B | x) = \int_A n(B|x, y) m\mathbb{F}_f(dy|x) \forall A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z} \tag{94}$$

In particular, $\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$ is a Borel measurable map $X \nrightarrow Y \times Z$. Thus equation 94 implies for all $A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z}$

$$\mathbb{P}^{\mathsf{YZ}|\mathsf{X}}_{\{\}}(\mathsf{Y}^{-1}(A)\cap B|x) = \int_A n(B|x,y)\mathbb{P}^{\mathsf{YZ}|\mathsf{X}}_{\{\}}\mathbb{F}_f(dy|x) \tag{95}$$

$$= \int_A n(B|x,y)\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\{\}}(dy|x) \tag{96}$$

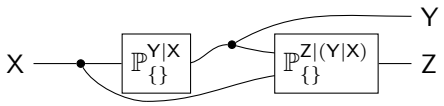
Where Equation 96 follows from Lemma 5.1.
Then, for any $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$

$$\mathbb{P}^{\mathsf{YZ}|\mathsf{X}}_{\{\}}(\mathsf{Y}^{-1}(A)\cap B|x) = \int_A n(B|x,y)\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\alpha}(dy|x) \tag{97}$$

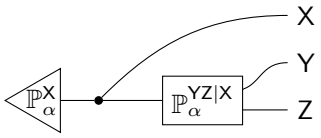
which implies n is a version of $\mathbb{P}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}_{\{\}}$. □

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* $\mathbb{P}^{\mathsf{Z}|\mathsf{YX}}_{\{\}}$

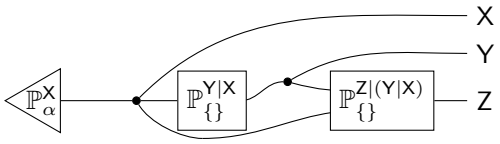
Proof. For arbitrary $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$



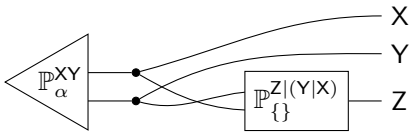
$$\mathbb{P}^{\mathsf{YZ}|\mathsf{X}}_{\alpha} = \tag{98}$$



(99)



(100)



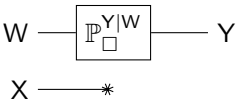
(101)

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 2.28, for each $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$ there exists $\mathbb{P}_\alpha^{\mathsf{Z}|\mathsf{Y}}$ such that

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{WX}} = \tag{102}$$



□

5.2 Extended conditional independence

Needs a support condition

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

Theorem 5.6. *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* $\mathbb{P}_\alpha^{\mathsf{VW}} := \mathbb{P}_\alpha^{\mathsf{W}} \odot \mathbb{P}^{\mathsf{V}|\mathsf{W}}$

$$\mathsf{A} \perp\!\!\!\perp_{\mathbb{P},ext} (\mathsf{B}, \mathsf{D})|(\mathsf{C}, \mathsf{E}) \iff \mathsf{A}^* \perp\!\!\!\perp_{\mathbb{P}_\alpha} (\mathsf{B}^*, \mathsf{D}^*)|(\mathsf{C}^*, \mathsf{E}^*) \tag{103}$$

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

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

Theorem 5.7. *Let* $\mathsf{A}^* = \mathsf{A} \circ \mathsf{V}$, $\mathsf{B}^* = \mathsf{B} \circ \mathsf{V}$, $\mathsf{C}^* = \mathsf{C} \circ \mathsf{V}$ *(*$(\mathsf{A}, \mathsf{B}, \mathsf{C})$ *are* $\mathcal{V}$*-measurable) and* $\mathsf{D}^* = \mathsf{D} \circ \mathsf{W}$,$\mathsf{E}^* = \mathsf{E} \circ \mathsf{W}$ *where* V, W *are discrete and* $\mathsf{W} = (\mathsf{D}^*, \mathsf{E}^*)$. *Then letting* $\mathbb{P}_\alpha^{\mathsf{VW}} := \mathbb{P}_\alpha^{\mathsf{W}} \odot \mathbb{P}^{\mathsf{V}|\mathsf{W}}$

$$\mathsf{A} \perp\!\!\!\perp^1_{\mathbb{P},ext} (\mathsf{B}, \mathsf{D})|(\mathsf{C}, \mathsf{E}) \iff \mathsf{A}^* \perp\!\!\!\perp_{\mathbb{P}} (\mathsf{B}^*, \mathsf{D}^*)|(\mathsf{C}^*, \mathsf{E}^*) \tag{104}$$

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

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 **??**, we have $\mathbb{P}^{\mathsf{A}|\mathsf{BCDE}}$ such that

$$\mathbb{P}_\beta^{\mathsf{A}^*\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} = \mathbb{P}_\beta^{\mathsf{D}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{B}^*\mathsf{C}^*|\mathsf{D}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{A}^*|\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} \tag{105}$$

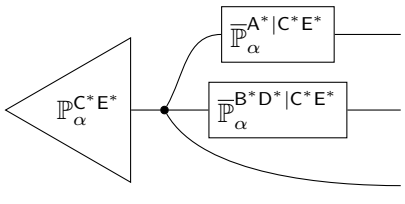
$$= \mathbb{P}_\beta^{\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{A}^*|\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} \tag{106}$$

$$= \mathbb{P}_\beta^{\mathsf{C}^*\mathsf{E}^*} \odot \mathbb{P}_\beta^{\mathsf{B}^*\mathsf{D}^*|\mathsf{C}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{A}^*|\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} \tag{107}$$

By Theorem 5.6, we have some α such that $\mathbb{P}_\alpha^{\mathsf{D}^*\mathsf{E}^*}$ is strictly positive on the range of $(\mathsf{D}^*, \mathsf{E}^*)$ and $\mathsf{A}^* \perp\!\!\!\perp_{\mathbb{P}_\alpha} (\mathsf{B}^*, \mathsf{D}^*)|(\mathsf{C}^*, \mathsf{E}^*)$.

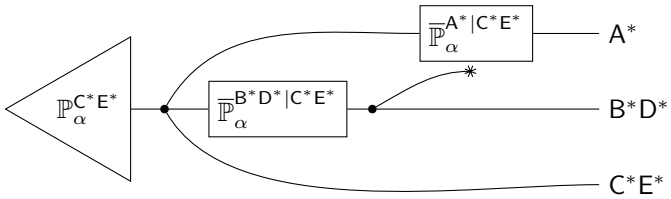
By independence, for some version of $\mathbb{P}^{\mathsf{A}|\mathsf{BCDE}}$:

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



(108)

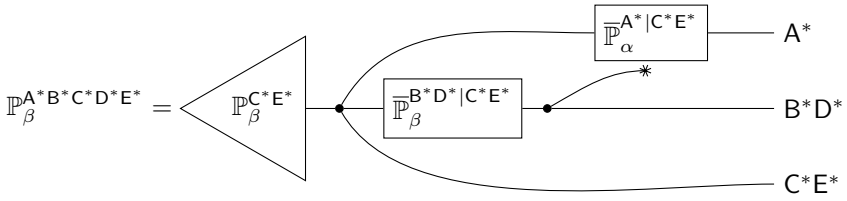
$=$



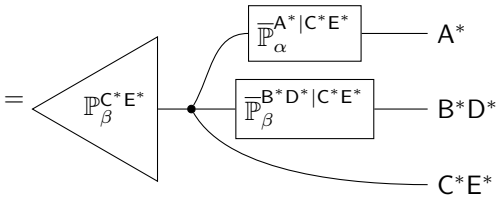
(109)

$$= \mathbb{P}_\alpha^{\mathsf{C}^*\mathsf{E}^*} \odot \mathbb{P}_\alpha^{\mathsf{B}^*\mathsf{D}^*|\mathsf{C}^*\mathsf{E}^*} \odot (\mathbb{P}_\alpha^{\mathsf{A}^*|\mathsf{C}^*\mathsf{E}^*} \otimes \mathrm{erase}_{BD}) \tag{110}$$

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



(111)

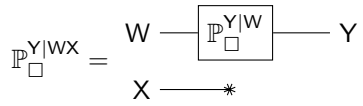


(112)

□

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

Definition 5.8 (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



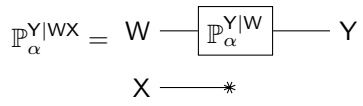
(113)

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

Definition 5.9 (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 5.10 (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

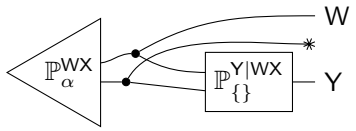


(114)

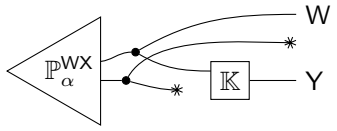
And so, by Theorem 2.28, $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{X}|\mathsf{W}$ for all $\alpha \in A$, and so $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{X}|\mathsf{W}$. Only if: For $\mathbb{P}_\alpha$ dominating $\mathbb{P}_{\{\}}$, by Theorem 2.28, there exists a version 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 5.11. *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 $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{X}|\mathsf{W}$.*

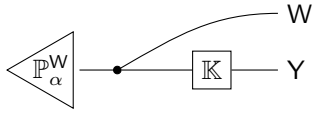
Proof. By Theorem 5.10, there is $\mathbb{K} : W \twoheadrightarrow Y$ such that for all α



$$\mathbb{P}_\alpha^{\mathsf{WY}} = \tag{115}$$



$$= \tag{116}$$



$$= \tag{117}$$

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.

5.3 Validity

Theorem 5.12 (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

$$\mu^{\mathsf{X}}(x) = \mu(\mathsf{X} \bowtie x) \tag{118}$$

$$= \sum_{x' \in X} \mathbb{J}(x')\mathbb{K}(\mathsf{X} \bowtie x | x') \tag{119}$$

$$= 0 \tag{120}$$

$$\neq \mathbb{J}(x) \tag{121}$$

□

Lemma 5.13 (Semidirect product is an intersection of probability sets)**.** *Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to (X, \mathcal{X})$, $\mathsf{Y} : \Omega \to (Y, \mathcal{Y})$, $\mathsf{Z} : \Omega \to (Z, \mathcal{Z})$ all standard measurable and valid candidate conditionals $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ and $\mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}$ defining probability sets $\mathbb{P}_{\{\}}$ and $\mathbb{Q}_{\{\}}$, then the probability set $\mathbb{R}_{\{\}}$ defined by $\mathbb{R}_{\{\}}^{\mathsf{YZ}|\mathsf{X}} := \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}$ is equal to $\mathbb{P}_{\{\}} \cap \mathbb{Q}_{\{\}}$.*

Proof. By assumption

$$\mathbb{R}_{\{\}}^{\mathsf{YZ}|\mathsf{X}} := \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{YX}} \tag{122}$$

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

$$\mathbb{R}_a^{\mathsf{XYZ}} = \mathbb{R}_a^{\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{YX}} \tag{123}$$

$$\implies \mathbb{R}_a^{\mathsf{XY}} = \mathbb{R}_a^{\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \tag{124}$$

$$\wedge \mathbb{R}_a^{\mathsf{XYZ}} = \mathbb{R}_a^{\mathsf{XY}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{YX}} \tag{125}$$

Thus $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ is a version of $\mathbb{R}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ and $\mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}$ is a version of $\mathbb{R}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}$ so $\mathbb{R}_{\{\}} \subset \mathbb{P}_{\{\}} \cap \mathbb{Q}_{\{\}}$.

Suppose there's an element $\mathbb{S}$ of $\mathbb{P}_{\{\}} \cap \mathbb{Q}_{\{\}}$ not in $\mathbb{R}_{\{\}}$. Then by definition of $\mathbb{R}_{\{\}}$, $\mathbb{R}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$ is not a version of $\mathbb{S}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$. But by construction of $\mathbb{S}$, $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ is a version of $\mathbb{S}^{\mathsf{Y}|\mathsf{X}}$ and $\mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}$ is a version of $\mathbb{S}^{\mathsf{Z}|\mathsf{YX}}$. But then by the definition of disintegration, $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}$ is a version of $\mathbb{S}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$ and so $\mathbb{R}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$ is a version of $\mathbb{S}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$, a contradiction. □

Lemma 5.14 (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

$$\forall A \in \mathcal{X} : \mathsf{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathsf{X}|*}(A|*) = 0 \tag{126}$$

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

$$\forall A \in \mathcal{X} : \mathsf{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathsf{X}}(A) = 0 \tag{127}$$

□

Lemma 5.15 (Copy-product of valid candidate conditionals is valid)**.** *Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ (all spaces standard measurable) and any valid candidate conditional $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ and $\mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}$, $\mathbb{P}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}$ is also a valid candidate conditional.*

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

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

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

$$\begin{aligned} \mathbb{R}^{\mathsf{YZ}|\mathsf{X}}(B \times C|x) &= \int_B \mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}(C|y,x)\mathbb{P}^{\mathsf{Y}|\mathsf{X}}(dy|x) && (128)\\ &\leq \mathbb{P}^{\mathsf{Y}|\mathsf{X}}(B|x) && (129)\\ &= 0 && (130) \end{aligned}$$

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

$$\begin{aligned} \mathbb{R}^{\mathsf{YZ}|\mathsf{X}}(B \times C|x) &= \int_B \mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}(C|y,x)\mathbb{P}^{\mathsf{Y}|\mathsf{X}}(dy|x) && (131)\\ &= 0 && (132) \end{aligned}$$

□

Corollary 5.16 (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 5.15 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 5.14. □

Theorem 5.17 (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 5.16.

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

$$(\mathbb{V}^{\mathsf{X}} \odot \mathbb{T}^{\mathsf{Y}|\mathsf{X}})(x, y) = \mathbb{T}^{\mathsf{Y}|\mathsf{X}}(y|x)\mathbb{V}^{\mathsf{X}}(x) \tag{133}$$

$$= \mathbb{T}^{\mathsf{Y}|\mathsf{X}}(y|x) \tag{134}$$

$$> 0 \tag{135}$$

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

Theorem 5.18 (Existence of valid conditional probabilities)**.** *Given a probability gap model $\mathbb{P}_\Box : A \to \Delta(\Omega)$ along with a valid conditional probability $\mathbb{P}_\Box^{\mathsf{XY}|\mathsf{W}}$, there exists a valid conditional probability $\mathbb{P}_\Box^{\mathsf{Y}|\mathsf{WX}}$.*

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

$$\mathbb{P}_\Box^{\mathsf{XY}|\mathsf{W}} = \mathbb{P}_\Box^{\mathsf{X}|\mathsf{W}} \odot \mathbb{P}_\Box^{\mathsf{Y}|\mathsf{WX}} \tag{136}$$

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

$$\mathbb{P}_\alpha^{\mathsf{WXY}} = \mathbb{P}_\alpha^{\mathsf{W}} \odot \mathbb{P}_\Box^{\mathsf{XY}|\mathsf{W}} \tag{137}$$

Thus

$$\mathbb{P}_\alpha^{\mathsf{WXY}} = \mathbb{P}_\alpha^{\mathsf{W}} \odot (\mathbb{P}_\Box^{\mathsf{X}|\mathsf{W}} \odot \mathbb{P}_\Box^{\mathsf{Y}|\mathsf{WX}}) \tag{138}$$

$$= (\mathbb{P}_\alpha^{\mathsf{W}} \odot \mathbb{P}_\Box^{\mathsf{X}|\mathsf{W}}) \odot \mathbb{P}_\Box^{\mathsf{Y}|\mathsf{WX}}) \tag{139}$$

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

$$(\mathsf{W}, \mathsf{X}) = (\text{idf}_{W \times X} \otimes \text{erasef}_Y) \circ (\mathsf{W}, \mathsf{X}, \mathsf{Y}) \tag{140}$$

By Lemma **??** together with Theorem 5.1 we have for all α:

$$\mathbb{P}_\alpha^{\mathsf{XW}} = \mathbb{P}_\alpha^{\mathsf{WXY}}(\text{id}_{W \times X} \otimes \text{erase}_Y) \tag{141}$$

$$= \mathbb{P}_\alpha^{\mathsf{W}} \odot (\mathbb{P}_\Box^{\mathsf{X}|\mathsf{W}} \odot \mathbb{P}_\Box^{\mathsf{Y}|\mathsf{WX}})(\text{id}_{W \times X} \otimes \text{erase}_Y) \tag{142}$$

$$= \mathbb{P}_\alpha^{\mathsf{W}} \odot \mathbb{P}_\Box^{\mathsf{X}|\mathsf{W}} \tag{143}$$

Then

$$\mathbb{P}_{\alpha}^{\mathsf{XWY}} = (\mathbb{P}_{\alpha}^{\mathsf{XW}}) \odot \mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{WX}}) \tag{144}$$

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

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

References

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`.

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.

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

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.

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.

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.

Alan Hájek. Interpretations of Probability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019. URL https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/.

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

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

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

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`.

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

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

Donald B. Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, March 2005. ISSN 0162-1459. doi: 10.1198/016214504000001880. URL `https://doi.org/10.1198/016214504000001880`.

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

Eyal Shahar. The association of body mass index with health outcomes: causal, inconsistent, or confounded? *American Journal of Epidemiology*, 170(8):957–958, October 2009. ISSN 1476-6256. doi: 10.1093/aje/kwp292.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, July 2002.

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

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

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

Appendix: