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

David Johnston

February 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 particular 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 models and action selection problems.

As a motivating example for our contribution, 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 effects" 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 real actions may instead be described by some more complicated variety of intervention operation. Even with this clarification, it appears that the relationship between interventions and actions is 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.

The question we focus on here is: when is there a well-defined causal effect of one variable on another? Many works on causal inference focus on the question of when we can *infer* the causal effect of one variable on another from a given sequence of data. In contrast, the question we focus on is not immediately appliccable to practical problems where investigators want to infer causal effects, but any such investigation must accept, implicitly at least, that causal effects do in fact exist. Thus we see our work as foundational to this key question of causation.

1.1 Our approach

We start with two attitudes (they're not precise enough to call assumptions):

- To understand "probabilistic causal effects", we need to study probabilistic models of decisions and consequences
- "Well-defined probabilistic causal effects" can be understood as symmetries of models of decisions and consequences

To the first proposition, one may object that counterfactual reasoning is the real theoretical foundation of causal modelling, and models of decisions and consequences are a special case of counterfactual reasoning (see Pearl and Mackenzie (2018) for example). To sidestep arguments of this nature: if all we accomplish is a better understanding of formal decision making and not causation as such, then our endeavour is still worthwhile.

The second proposition is similar to De Finetti's analysis of the concept of a sequence of events distributed according to a "constant but unknown probability $\mathbb{Q}$". De Finetti observed that, while one may use a probability model $\mathbb{P}$ to express an uncertain forecast of the outcomes of a sequence of events, the probability $\mathbb{Q}$

itself seems to represent something of a different kind. In particular, interpreting $\mathbb{Q}$ requires some additional theory of what it means for a probability model to be correct, while interpreting $\mathbb{P}$ only requires us to say what it means for an outcome to be realised. De Finetti's solution to this question was to propose that an unknown probability $\mathbb{Q}$ could be understood as a feature of a forecast $\mathbb{P}$ which has the property of exchangeability.

In a similar fashion, we observe that one can use a probabilistic model to help make a decision without any theory of what it means for some variable to have a causal effect on some other variable. Thus, like the constant but uknown probability $\mathbb{Q}$, a "fixed but uknown causal effect $\mathsf{Q}(\mathsf{Y}|do(\mathsf{X}))$" requires a theory of what it means for a causal effect to be correct in addition to a probabilistic model of the consequences of decisions. By analogy with De Finetti's reasoning, we propose a theory of causal effects as properties of probabilistic decision models that have a certain type of symmetry that we call *response contractibility*.

As we have just mentioned, we aren't proposing that this is a compelling account of "causal effects" in every sense in which the phrase is ever used. However, many causal investigations involve analysing sequences of events that are in some sense repeatable with the aim of helping people interested in influencing similar events in the future to make good decisions. Our theory applies to analysis in this setting. We are studying a particular kind of causal effect which we call a *repeatable response*. Thus, our motivating question is more precisely stated as "when do probabilistic decision models entail the existence of fixed but unknown conditional probabilities representing repeatable responses?"

To answer this question, we introduce two different pieces of theory. Firstly, we present a mathematical theory of *probability sets*, which extends the standard theory of probability by replacing individual probability measures with sets of probability measures. This extension allows us to model situations in which:

- We are able to decide on one choice from a number of different possible choices
- The result of each decision is associated with a different probability measure
- There are some features of the resulting probability measures that are common to every choice available

We note that there are similarities between the theory of probability sets and *imprecise probability* (Walley, 1991), but the precise connections between our theory and different theories of imprecise probability are an open question.

We use the theory of probability sets reason about models of decision problems. However, reasoning about a given model of a problem is only half the story – we also need to be able to decide when a model is appropriate for a problem. This motivates the second piece of theory presented here: a theory of variables and measurement procedures. This theory is somewhat vague, and we don't see a way to avoid vagueness. We propose *measurement procedures* that are function-like things whose "domain" is what we vaguely refer to as "the real

world". Executing a measurement procedure involves interacting with the real world somehow such that, ultimately, a unique element of a well-defined mathematical set is returned. Because measurement procedures have mathematical sets as their "codomain", they can be composed with functions. Because their "domain" is the real world, we cannot compose functions with measurement procedures. Variables are functions – with well-defined domains and codomains – that we identify with measurement procedures.

This theory is suggested by many introductions to probability theory. For example, Boole (1862) discusses elements of "the actual problem", described in natural language, and a corresponding collection of "ideal events" which models the actual problem and also also obey postulates of probability theory. Feller (1968) describes experiments and observations as "things whose results take unique values in well-defined mathematical sets". However, our theory is most informed by the theory of random variables presented by Menger (2003), whom we credit with many of the insights, although our terminology and notation differs somewhat.

1.2 Contributions

A secondary contribution of this paper is the notion of *validity* of a model represented by a probability set. This is simply the requirement that the probability set is nonempty. The problem of whether a probability set is valid is analogous to the problem of whether a probability distribution satisfying a collection of constraints exists, discussed in Vorobev (1962), although our use of the concept is somewhat more elementary. We discuss how an incautious attempt to build a model of "interventions on body mass index" can yield an invalid model.

There are two main contributions. The first is a formal result akin to De Finetti's representation theorem (de Finetti, [1937] 1992). De Finetti's theorem shows that *exchangeability* of a probability model is equivalent – in a certain sense – to the existence of a "fixed but unknown" probability distribution over a sequence of observations. We introduce a symmetry called *causal contractibility* and show that it is – in a similar sense – equivalent to the existence of a "fixed but unknown" conditional probability representing the response of one variable to the value of another.

There's a logic issue I still need to work out regarding whether causal contractibility assumes determinism or not

Our second contribution is to consider what kinds of measurement processes support a judgement of causal contractibility. We show that subtly different descriptions of measurement process can support or fail to support such a judgement. In particular, we examine how judgements of causal contractibility might be supported when a decision deterministically fixes a sequence of choices at a point in time when they all look equivalent to a decision maker, but not supported by a measurement process that is described identically except the choices are not deterministically fixed. We also discuss how causal contractibility for nondeterministic variables can follow from a prior judgement of causal

contractibility in combination with a certain kind of conditional independence that we call *proxy control*.

We consider it an open question whether judgements of causal contractibility are supported by any measurement procedure that isn't described either of the options we consider – that is, by measurement procedures that don't involve deterministically selecting choices from a position of "epistemic indifference" or from proxy control in combination with a prior judgement of causal contractibility.

2 Variables in probabilistic models

Our main question concerns the existence of causal relationships between *variables*. If we want to offer a clear account of what this means, we need to start with a clear account of what variables are. Both observed an unobserved variables play important roles in causal modelling and we think it is worth clarifying what variables of either type refer to. We will start with observed variables, which we consider to be parts of our model whose role is to "point to the parts of the world the model is explaining". Unobserved variables, on the other hand, are parts of the model that do not refer to the external world but may be introduced, for example, for notational convenience.

Our approach in short is: a probabilistic model is associated with a particular experiment or measurement procedure. The measurment procedure yields values in a well-defined set. Observable results are obtained by applying well-defined functions to the result of this procedure. The observable sample space is the set of values that can be obtained from the experiment, and observable variables are the functions associated with particular observable results. We extend the set of values obtained from the observable sample space to a sample space that can contains both obserable and unobservable variables. Unobservable variables, like observable variables, are functions on the sample space, but they do not correspond to any observable results.

As far as we know, distinguishing variables from procedures is somewhat non-standard, but we feel it is useful to distinguish the formal elements of the theory (variables) from the semi-formal elements (measurement procedures). Both variables and procedures are often discussed in statistical texts. For example, Pearl (2009) offers the following two, purportedly equivalent, definitions of variables:

> By a *variable* we will mean an attribute, measurement or inquiry that may take on one of several possible outcomes, or values, from a specified domain. If we have beliefs (i.e., probabilities) attached to the possible values that a variable may attain, we will call that variable a random variable.
>
> This is a minor generalization of the textbook definition, according to which a random variable is a mapping from the fundamental probability set (e.g., the set of elementary events) to the real line. In our

> definition, the mapping is from the fundamental probability set to any set of objects called "values," which may or may not be ordered.

Our view is that the first definition is a definition of a procedure, while the second is a definition of a variable. Variables model procedures, but they are not the same thing. We can establish this by noting that, under our definition, every procedure of interest – that is, all procedures that can be written $f \circ \mathcal{S}$ for some f – is modeled by a variable, but there may be variables defined on Ω that do not factorise through $\mathcal{S}$, and these variables do not model procedures.

We illustrate this approach with the example of Newton's second law in the form $\mathsf{F} = \mathsf{MA}$. This model relates "variables" F, M and A. As Feynman (1979) noted, in order to understand this law, we must bring some pre-existing understanding of force, mass and acceleration independent of the law itself. Furthermore, we contend, this knowledge cannot be expressed in any purely mathematical statement. In order to say wha the net force on a given object is, even a highly knowledgeable physicist will have to go and do some measurements, which is a procedure that they carry out involving interacting with the real world somehow and obtaining as a result a vector representing the net forces on that object.

That is, the variables F, M and A are referring to the *results of measurement procedures*. We will introduce a separate notation to refer to these measurement procedures – $\mathcal{F}$ is the procedure for measuring force, $\mathcal{M}$ and $\mathcal{A}$ for mass and acceleration respectively. A measurement procedure $\mathcal{F}$ is akin to Menger (2003)'s notion of variables as "consistent classes of quantities" that consist of pairing between real-world objects and quantities of some type. Force $\mathcal{F}$ itself is not a well-defined mathematical thing, as measurement procedures are not mathematically well-defined. At the same time, the set of values it may yield *are* well-defined mathematical things. No actual procedure can be guaranteed to return elements of a mathematical set known in advance – anything can fail – but we assume that we can study procedures reliable enough that we don't lose much by making this assumption.

Note that, because $\mathcal{F}$ is not a purely mathematical thing, we cannot perform mathematical reasoning with $\mathcal{F}$ directly. Rather, we introduce a variable F which, as we will see, is a well-defined mathematical object, assert that it corresponds to $\mathcal{F}$ and conduct our reasoning using F.

2.1 Measurment procedures

Definition 2.1 (Measurement procedure)**.** A *measurement procedure* $\mathcal{B}$ is a procedure that involves interacting with the real world somehow and delivering an element of a mathematical set X as a result. A procedure is given the font $\mathcal{B}$, we say it takes values in X.

Definition 2.2 (Values yielded by procedures)**.** $\mathcal{B} \bowtie x$ is the proposition that the the procedure $\mathcal{B}$ will yield the value $x \in X$. $\mathcal{B} \bowtie A$ for $A \subset X$ is the proposition $\vee_{x \in A} \mathcal{B} \bowtie x$.

Definition 2.3 (Equivalence of procedures)**.** Two procedures $\mathcal{B}$ and $\mathcal{C}$ are equal if they both take values in X and $\mathcal{B} \bowtie x \iff \mathcal{C} \bowtie x$ for all $x \in X$. If they involve different measurment actions in the real world but still necessarily yield the same result, we say they are equal.

It is worth noting that this notion of equivalence identifies procedures with different real-world actions. For example, "measure the force" and "measure everything, then discard everything but the force" are often different – in particular, it might be possible to measure the force only before one has measured everything else. Thus the result yielded by the first procedure could be available before the result of the second. However, if the first is carried out in the course of carrying out the second, they both yield the same result in the end and so we treat them as equivalent.

Measurement procedures are like functions without well-defined domains. Just like we can compose functions with other functions to create new functions, we can compose measurement procedures with functions to produce new measurement procedures.

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

For example, $\mathcal{MA}$ is the composition of $h : (x, y) \mapsto xy$ with the procedure $(\mathcal{M}, \mathcal{A})$ that yields the mass and acceleration of the same object. Measurement procedure composition is associative:

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

One might whether there is also some kind of "append" operation that takes a standalone $\mathcal{M}$ and a standalone $\mathcal{A}$ and returns a procedure $(\mathcal{M}, \mathcal{A})$. Unlike function composition, this would be an operation that acts on two procedures rather than a procedure and a function. Thus this "append" combines real-world operations somehow, which might introduce additional requirements (we can't just measure mass and acceleration; we need to measure the mass and acceleration of the same object at the same time), and may be under-specified. For example, measuring a subatomic particle's position and momentum can be done separately, but if we wish to combine the two procedures then the we can get different results depending on the order in which we combine them.

Our approach here is to suppose that there is some complete measurement procedure $\mathcal{S}$ to be modeled, which takes values in the observable sample space

$(\Psi, \mathcal{E})$ and for all measurement procedures of interest there is some f such that the procedure is equivalent to $f \circ \mathcal{S}$ for some f. In this manner, we assume that any problems that arise from a need to combine real world actions have already been solved in the course of defining $\mathcal{S}$.

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

In probability theory, another standard kind of measurable space considered is isomorphic to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, i.e. the reals with the Borel sigma-algebra. It is not obvious to us why this would be a natural choice to represent possible results of an actual measurement. It is possible that a Borel measurable space is an appropriate idealisation of "floating point" measurements, but we don't have a precise argument for this.

2.2 Observable variables

Our total procedure $\mathcal{S}$ represents a large collection of subprocedures of interest, each of which can be obtained by composition of some function with $\mathcal{S}$. We call the pair consisting of a subprocedure of interest $\mathcal{X}$ along with the variable X used to obtain it from $\mathcal{S}$ an *observable variable*.

Definition 2.5 (Observable variable)**.** Given a measurement procedure $\mathcal{S}$ taking values in $(\Psi, \mathcal{E})$, an observable variable is a pair $(\mathsf{X} \circ \mathcal{S}, \mathsf{X})$ where $\mathsf{X} : (\Psi, \mathcal{E}) \to (X, \mathcal{X})$ is a measurable function and $\mathcal{X} := \mathsf{X} \circ \mathcal{S}$ is the measurement procedure induced by X and $\mathcal{S}$.

For the model $\mathsf{F} = \mathsf{MA}$, for example, suppose we have a total measurement procedure $\mathcal{S}$ that yields a triple (force, mass, acceleration) taking values in the sets X, Y, Z respectively. Then we can define the "force" variable $(\mathcal{F}, \mathsf{F})$ where $\mathcal{F} := \mathsf{F} \circ \mathcal{S}$ and $\mathsf{F} : X \times Y \times Z \to X$ is the projection function onto X.

A measurement procedure yields a particular value when it is completed. We will call a proposition of the form "$\mathcal{X}$ yields x" an *observation*. Note that $\mathcal{X}$ need not be a total procedure here. Given the total procedure $\mathcal{S}$, a variable $\mathsf{X} : \Psi \to X$ and the corresponding procedure $\mathcal{X} = \mathsf{X} \circ \mathcal{S}$, the proposition "$\mathcal{X}$ yields x" is equivalent to the proposition "$\mathcal{S}$ yields a value in $\mathsf{X}^{-1}(x)$". Because of this, we define the *event* $\mathsf{X} \bowtie x$ to be the set $\mathsf{X}^{-1}(x)$.

Definition 2.6 (Event)**.** Given the total procedure $\mathcal{S}$ taking values in Ψ and an observable variable $(\mathsf{X} \circ \mathcal{S}, \mathsf{X})$ for $\mathsf{X} : \Psi \to X$, the *event* $\mathsf{X} \bowtie x$ is the set $\mathsf{X}^{-1}(x)$ for any $x \in X$.

If we are given an observation "$\mathcal{X}$ yields x", then the corresponding event $\mathsf{X} \bowtie x$ is *compatible with this observation*.

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

An *impossible event* is the empty set. If $\mathsf{X} \bowtie x = \emptyset$ this means that we have identified no possible outcomes of the measurement process $\mathcal{S}$ compaible with the observatoin "$\mathcal{X}$ yields x".

2.3 Model variables

Observable variables are special in the sense that they are tied to a particular measurement procedure $\mathcal{S}$. However, the measurment procedure $\mathcal{S}$ does not enter into our mathematical reasoning; it guides our construction of a mathematical model, but once this is done mathematical reasoning proceeds entirely with mathematical objects like sets and functions, with no further reference to the measurement procedure.

A *model variable* is what we are left with if we take an observable variable and discard most of the total measurement procedure $\mathcal{S}$, retaining only its set of possible values $(\Psi, \mathcal{E})$. A model variable is simply a measurable function with domain Ψ.

Model variables do not have to be derived from observable variables. We may instead choose a sample space for our model $(\Omega, \mathcal{F})$ that does not correspond to the possible values that $\mathcal{S}$ might yield. In that case, we require a surjective model variable $\mathsf{S} : \Omega \to \Psi$, and every observable variable $(\mathsf{X}' \circ \mathcal{S}, \mathsf{X}')$ is associated with the model variable $\mathsf{X} := \mathsf{X}' \circ \mathsf{S}$.

2.4 Variable sequences

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

2.5 Actions

We also deal with variables that represent "decisions" or "actions". If we consider what happens in the real world, there's a difference between an action and a measurement in that the former involves "doing stuff" while the latter involves just "seeing stuff". However, for the purposes of mathematically modelling actions and measurements, we note that any procedure for measuring the choice of action should always agree with the action chosen. Thus, by Definition 2.3, a procedure for choosing an action is equivalent to a procedure for "measuring" which action was chosen. In particular, if I have a functional rule for choosing an action, then I have a measurement procedure. For example, if I take a measurement $\mathcal{X}$ and then pick an action by applying the function f to the result, I then have a measurement procedure $\mathcal{D} = f \circ \mathcal{X}$ for the action I chose.

3 Probability sets

3.1 The roles of variables and probabilistic models

The sample space $(\Omega, \mathcal{F})$ along with the measurement procedure $\mathcal{S}$ and the associated model variable S is a "model skeleton". The criterion of *compatibility with observation* establishes a relation between observations and elements of $\mathcal{F}$.

The basic kind of problem we want to consider is one in which we wish to decide upon an action that we expect will yield good consequences. We suppose that whether a consequence is good or not can somehow be deduced from the result of $\mathcal{S}$. However, we do not know the result of $\mathcal{S}$, so we need to say something about the result we expect to see for each action we could choose.

It is common to to represent uncertain knowledge about the outcomes of not-yet-performed measurements using probabilistic models, and we follow this well-trodden path. However, we do need to generalise common practice somewhat, because we need a model that tells us that different consequences may arise from deciding on different actions.

We use probability sets and probability gap models to represent decision problems. A probability set is a set of probability measures on a common sample space $(\Omega, \mathcal{F})$, and a probability gap model is a probability set along with a collection of subsets (the terminology comes from Hájek (2003)). A decision problem presents us with a set of choices, and we assume that each choice is associated with a probability set representing uncertain knowledge (or best guesses) about the outcome of this choice. A probability gap model is the collection of all probability sets associated with a choice, along with the union of all of these sets. The union of all of the individual choice sets represents what we know about the outcome regardless of which choice is decided on.

Our use of probability sets to represent uncertain knowledge about the outcome of each choice is not the result of a strong opinion that probability sets are the best way to do this. We've already had to introduce probability sets to handle different choices in the first place and we don't see any harm in continuing to use them for this additional purpose. A model in which a unique probability distribution is associated with each choice is simply a special case of this setup, where the probability set associated with each choice is of size 1.

A great deal of standard probability theory is appliccable to teasoning with probability sets, and readers may be quite familiar with much of this. In particular, our notions of uniform conditional probability and uniform conditional independence are similar in many ways to the familiar notions of conditional probability and conditional indpendence, with the different being that – even in finite sets – the former do not always exist. We also make use of a diagrammatic notation for Markov kernels (or stochastic functions) taken from the categorical study of probability theory, which may be less familiar.

3.2 Standard probability theory

Definition 3.1 (Measurable space)**.** A measurable space $(X, \mathcal{X})$ is a set X along with a σ-algebra of subsets $\mathcal{X}$.

We use a number of shorthands for measurable spaces:

- Where the choice of σ-algebra is unambiguous, we will just use the set name X to refer to X along with a σ-algebra $\mathcal{X}$
- For a discrete set X, the sigma-algebra $\mathcal{X}$ referred to with the same letter is the discrete sigma-algebra
- For a continuous set X, the sigma-algebra $\mathcal{X}$ referred to with the same letter is the Borel sigma-algebra

Definition 3.2 (Probability measure)**.** Given a measurable 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 3.3 (Markov kernel)**.** Given measurable spaces $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$, a Markov kernel $\mathbb{Q} : X \twoheadrightarrow Y$ is a map $Y \times \mathcal{X} \to [0, 1]$ such that

1. $y \mapsto \mathbb{Q}(A|y)$ is $\mathcal{Y}$-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 3.4 (Delta measure)**.** Given a measureable space $(X, \mathcal{X})$ and $x \in X$, $\delta_x \in \Delta(X)$ is the measure defined by $\delta_x(A) := [\![x \in A]\!]$ for all $A \in \mathcal{X}$

Definition 3.5 (Probability space)**.** A probability space is a triple $(\mu, \Omega, \mathcal{F})$, where μ is a base measure on $\mathcal{F}$ and $(\Omega, \mathcal{F})$ is a measurable space.

Definition 3.6 (Variable)**.** Given a measureable space $(\Omega, \mathcal{F})$ and a measurable space of values $(X, \mathcal{X})$, an *X-valued variable* is a measurable function $\mathsf{X} : (\Omega, \mathcal{F}) \to (X, \mathcal{X})$.

Definition 3.7 (Sequence of variables)**.** Given a measureable space $(\Omega, \mathcal{F})$ and two variables $\mathsf{X} : (\Omega, \mathcal{F}) \to (X, \mathcal{X})$, $\mathsf{Y} : (\Omega, \mathcal{F}) \to (Y, \mathcal{Y})$, $(\mathsf{X}, \mathsf{Y}) : \Omega \to X \times Y$ is the variable $\omega \mapsto (\mathsf{X}(\omega), \mathsf{Y}(\omega))$.

Definition 3.8 (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 $\mu^{\mathsf{X}}(A) := \mu(\mathsf{X}^{-1}(A))$ for any $A \in \mathcal{X}$.

Definition 3.9 (Distribution-kernel products)**.** Given $(X, \mathcal{X})$, $(Y, \mathcal{Y})$ a probability distribution $\mu \in \Delta(X)$ and a Markov kernel $\mathbb{K} : X \twoheadrightarrow Y$, $\mu\mathbb{K}$ is a probability distribution on $(Y, \mathcal{Y})$ defined by

$$\mu\mathbb{K}(A) := \int_X \mathbb{K}(A|x)\mu(\mathrm{d}x) \tag{5}$$

for all $A \in \mathcal{Y}$.

Definition 3.10 (Kernel-kernel products)**.** Given $(X, \mathcal{X})$, $(Y, \mathcal{Y})$, $(Z, \mathcal{Z})$ and Markov kernels $\mathbb{K} : X \nrightarrow Y$ and $\mathbb{L} : Y \nrightarrow Z$, $\mathbb{KL}$ is a Markov kernel $X \nrightarrow Z$ defined by

$$\mathbb{KL}(A|x) := \int_Y \mathbb{L}(A|y)\mathbb{K}(\mathrm{d}y|x) \tag{6}$$

for all $A \in \mathcal{Z}$.

Lemma 3.11 (Marginal distribution as a kernel product)**.** *Given a probability space $(\mu, \Omega, \mathcal{F})$ and a variable $\mathsf{X} : \Omega \to (X, \mathcal{X})$, define $\mathbb{F}_{\mathsf{X}} : \Omega \nrightarrow X$ by $\mathbb{F}_{\mathsf{X}}(A|\omega) = \delta_{\mathsf{X}(\omega)}(A)$, then*

$$\mu^{\mathsf{X}} = \mu\mathbb{F}_{\mathsf{X}} \tag{7}$$

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

$$\begin{aligned} \mu\mathbb{F}_{\mathsf{X}}(A) &= \int_\Omega \delta_{\mathsf{X}(\omega)}(A)\mathrm{d}\mu(\omega) && (8)\\ &= \int_{\mathsf{X}^{-1}(\omega)} \mathrm{d}\mu(\omega) && (9)\\ &= \mu^{\mathsf{X}}(A) && (10) \end{aligned}$$

□

3.3 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 $\{*\}$. This will prove useful in further developments, as it means that we can treat probability distributions and Markov kernels as different varieties of the same kind of thing.

Definition 3.12 (Probability measures as Markov kernels)**.** Given a measurable space $(X, \mathcal{X})$ and $\mu \in \Delta(X)$, the Markov kernel $\mathbb{K} : \{*\} \nrightarrow X$ associated with μ is given by $\mathbb{K}(A|*) = \mu(A)$ for all $A \in \mathcal{X}$.

We will use probability measures and their associated Markov kernels interchangeably, as it is transparent how to get from one to another.

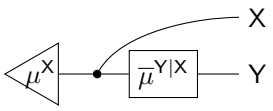
Conditional probability distributions are "Markov kernel annotated with variables".

Definition 3.13 (Conditional distribution)**.** Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, the probability of Y given X is any Markov

kernel $\mu^{\mathsf{Y}|\mathsf{X}} : X \twoheadrightarrow Y$ such that

$$\mu^{\mathsf{XY}}(A \times B) = \int_A \mu^{\mathsf{Y}|\mathsf{X}}(B|x)\mathrm{d}\mu^{\mathsf{X}}(x) \qquad \forall A \in \mathcal{X}, B \in \mathcal{Y} \tag{11}$$

$$\iff \tag{12}$$

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

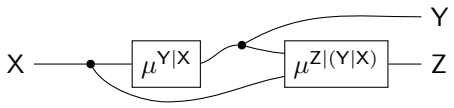
(13)

We define higher order conditionals as "conditionals of conditionals".

Definition 3.14 (Higher order conditionals)**.** Given a probability space $(\mu, \Omega, \mathcal{F})$ 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 $\mu^{\mathsf{Y}|\mathsf{X}}$,

$$\mu^{\mathsf{ZY}|\mathsf{X}}(B \times C|x) = \int_B \mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}(C|x, y)\mu^{\mathsf{Y}|\mathsf{X}}(dy|x) \tag{14}$$

$$\iff \tag{15}$$



$$\mu^{\mathsf{ZY}|\mathsf{X}} = \tag{16}$$

Higher order conditionals are useful because $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ is a version of $\mu^{\mathsf{Z}|\mathsf{YX}}$, so if we're given $\mu^{\mathsf{ZY}|\mathsf{X}}$ but not μ itself, we use the higher order conditional $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ as a version of $\mu^{\mathsf{X}|\mathsf{YX}}$. This also hold for conditional with respect to probability sets, which we will introduce later (Theorem 6.4).

Furthermore, given $\mu^{\mathsf{XY}|\mathsf{Z}}$ and X, Y standard measurable, it has recently been proven that a higher order conditional $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ exists Bogachev and Malofeev (2020), Theorem 3.5. See also Theorem 6.3 for the extension of this theorem to probability sets.

3.4 Probability sets

I've accepted Bob's comments about the notation, but I haven't actually changed the notation at this point

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

Given a probability set $\mathbb{P}_{\{\}}$, we define marginal and conditional probabilities as probability measures and Markov kernels that satisfy Definitions 3.8 and 3.13 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 3.21) with respect to probability sets.

Definition 3.16 (Marginal probability with respect to a probability set)**.** Given a sample space $(\Omega, \mathcal{F})$, a variable $\mathsf{X} : \Omega \to X$ and a probability set $\mathbb{P}_{\{\}}$, the marginal distribution $\mathbb{P}^{\mathsf{X}}_{\{\}} = \mathbb{P}^{\mathsf{X}}_{\alpha}$ for any $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$ if a distribution satisfying this condition exists. Otherwise, it is undefined.

Definition 3.17 (Uniform conditional distribution with respect to a probability set)**.** Given a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$ and a probability set $\mathbb{P}_{\{\}}$, a uniform conditional distribution $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\{\}}$ is any Markov kernel $X \twoheadrightarrow Y$ such that $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\{\}}$ is an $\mathsf{Y}|\mathsf{X}$ conditional probability of $\mathbb{P}_\alpha$ for all $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$. If no such Markov kernel exists, $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\{\}}$ is undefined.

Definition 3.18 (Uniform higher order conditional distribuition with respect to a probability set)**.** Given a sample space $(\Omega, \mathcal{F})$, 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 uniform higher order conditional $\mathbb{P}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}_{\{\}}$ is any Markov kernel $X \times Y \twoheadrightarrow Z$ that is a higher order conditional of some version of $\mathbb{P}^{\mathsf{ZY}|\mathsf{X}}_{\{\}}$. If no $\mathbb{P}^{\mathsf{ZY}|\mathsf{X}}_{\{\}}$ exists, $\mathbb{P}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}_{\{\}}$ is undefined.

Under the assumption of standard measurable spaces, the existence of a uniform conditional distribution $\mathbb{P}^{\mathsf{ZY}|\mathsf{X}}_{\{\}}$ implies the existence of a higher order conditional $\mathbb{P}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}_{\{\}}$ with respect to the same probability set (Theorem 6.3). $\mathbb{P}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}_{\{\}}$ is in turn a version of the uniform conditional distribution $\mathbb{P}^{\mathsf{Z}|\mathsf{YX}}_{\{\}}$ (Theorem 6.4). Thus, from the existence of $\mathbb{P}^{\mathsf{ZY}|\mathsf{X}}_{\{\}}$ we can derive the existence of $\mathbb{P}^{\mathsf{Z}|\mathsf{YX}}_{\{\}}$.

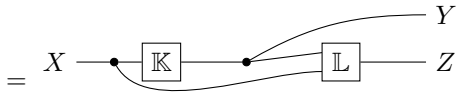
3.5 Semidirect product and almost sure equality

The operation used in Equation 13 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 using $\odot$: $\mathbb{K} \overset{\mu^{\mathsf{X}}}{\cong} \mathbb{L}$ if $\mu^{\mathsf{X}} \odot \mathbb{K} = \mu^{\mathsf{X}} \odot \mathbb{L}$ (note that this latter equality is strict; both semidirect products must assign the same measure to the same measurable sets). Thus if two terms are almost surely equal, they are substitutable when they both appear in a semidirect product.

Definition 3.19 (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

$$\mathbb{K} \odot \mathbb{L} := \mathrm{copy}_X(\mathbb{K} \otimes \mathrm{id}_X)(\mathrm{copy}_Y \otimes \mathrm{id}_X)(\mathrm{id}_Y \otimes \mathbb{L}) \tag{17}$$

$$= \tag{18}$$



$$\iff \tag{19}$$

$$(\mathbb{K} \odot \mathbb{L})(A \times B|x) = \int_A \mathbb{L}(B|y,x)\mathbb{K}(dy|x) \qquad A \in \mathcal{Y}, B \in \mathcal{Z} \tag{20}$$

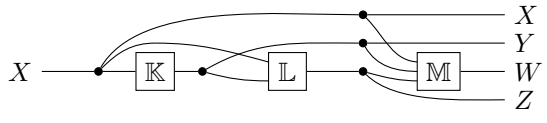
Lemma 3.20 (Semidirect product is associative)**.** *Given* $\mathbb{K} : X \nrightarrow Y$*,* $\mathbb{L} : Y \times X \nrightarrow Z$ *and* $\mathbb{M} : Z \times Y \times X \nrightarrow W$

$$(\mathbb{K} \odot \mathbb{L}) \odot \mathbb{Z} = \mathbb{K} \odot (\mathbb{L} \odot \mathbb{Z}) \tag{21}$$

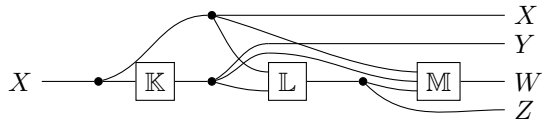
$$\tag{22}$$

Proof.

$$(\mathbb{K} \odot \mathbb{L}) \odot \mathbb{M} = \tag{23}$$



$$= \tag{24}$$



$$= \mathbb{K} \odot (\mathbb{L} \odot \mathbb{M}) \tag{25}$$

□

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

Definition 3.21 (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 $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$,

$$\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{K} = \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{L} \tag{26}$$

Lemma 3.22 (Uniform conditional distributions are almost surely equal)**.** *If* $\mathbb{K} : X \nrightarrow Y$ *and* $\mathbb{L} : X \nrightarrow Y$ *are both versions of* $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ *then* $\mathbb{K} \overset{\mathbb{P}_{\{\}}}{\cong} \mathbb{L}$

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

$$\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{K} = \mathbb{P}_\alpha^{\mathsf{XY}} \tag{27}$$

$$= \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{L} \tag{28}$$

□

Lemma 3.23 (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* $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{K} \overset{\mathbb{P}_{\{\}}}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{L} \tag{29}$$

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

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

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

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

□

Lemma 3.24 (Semidirect product of uniform conditional distributions is a joint uniform conditional distribution)**.** *Given a probability set* $\mathbb{P}_{\{\}}$ *on* $(\Omega, \mathcal{F})$*, variables* $\mathsf{X} : \Omega \to X$*,* $\mathsf{Y} : \Omega \to Y$ *and uniform conditional distributions* $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ *and* $\mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{XY}}$*, then* $\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$ *exists and is equal to*

$$\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}} = \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{XY}} \tag{33}$$

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

$$\mathbb{P}_\alpha^{\mathsf{XYZ}} = \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_\alpha^{\mathsf{YZ}|\mathsf{X}} \tag{34}$$

$$= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_\alpha^{\mathsf{Z}|\mathsf{YX}}) \tag{35}$$

$$= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}) \tag{36}$$

□

3.6 Maximal probability sets and valid conditionals

So far we have defined probability sets and conditional probabilities as Markov kernels that can sometimes be derived from a probability set. For the purposes of analysing decision models, we are often interested in working in the opposite direction: starting with conditional probabilities and working with probability sets defined by them. We call these *maximal probability sets*.

We need to be a little bit careful when we proceed in this fashion: we can't take an arbitrary Markov kernel $\kappa : X \twoheadrightarrow Y$ and declare it to be a conditional probability $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ for some $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$ and a maximal probability set $\mathbb{P}_{\{\}}$. The reason for this is that some collections of variables cannot have arbitrary conditional probabilities, and so $\mathbb{P}_{\{\}}$ may in fact be the empty set. We address this with the notion of validity; a *valid distribution* is a distribution associated with a particular variable that defines a nonempty set of base measures on Ω (Theorem 6.9), and *valid conditionals* are a set of conditional probabilities closed under $\odot$ and reducing to valid distributions when conditioning on a trivial variable (Lemma 6.12).

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

Definition 3.25 (Valid distribution)**.** Given $(\Omega, \mathcal{F})$ and a variable $\mathsf{X} : \Omega \to X$, an X-valid probability distribution is any probability measure $\mathbb{K} \in \Delta(X)$ such that $\mathsf{X}^{-1}(A) = \emptyset \implies \mathbb{K}(A) = 0$ for all $A \in \mathcal{X}$.

Definition 3.26 (Valid conditional)**.** Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ a $\mathsf{Y}|\mathsf{X}$-*valid conditional probability* is a Markov kernel $\mathbb{L} : X \twoheadrightarrow Y$ that assigns probability 0 to impossible events, unless the argument itself corresponds to an impossible event:

$$\forall B \in \mathcal{Y}, x \in X : (\mathsf{X}, \mathsf{Y}) \bowtie \{x\} \times B = \emptyset \implies (\mathbb{L}(B|x) = 0) \vee (\mathsf{X} \bowtie \{x\} = \emptyset) \tag{37}$$

Definition 3.27 (Maximal probability set)**.** Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and a $\mathsf{Y}|\mathsf{X}$-valid conditional probability $\mathbb{L} : X \twoheadrightarrow Y$ the maximal probability set $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}[M]}$ associated with $\mathbb{L}$ is the probability set such that for all $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$, $\mathbb{L}$ is a version of $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}$.

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

Lemma 6.12 shows that the semidirect product of any pair of valid conditoinal probabilities is itself a valid conditional. Suppose we have some collection of $\mathsf{X}_i|\mathsf{X}_{[i-1]}$-valid conditionals $\{\mathbb{P}_i^{\mathsf{X}_i|\mathsf{X}_{[i-1]}} | i \in [n]\}$; then recursively taking the semidirect product $\mathbb{M} := \mathbb{P}_1^{\mathsf{X}_1} \odot (\mathbb{P}_2^{\mathsf{X}_2|\mathsf{X}_1} \odot ...)$ yields a $\mathsf{X}_{[n]}$ valid distribution. Furthermore, the maximal probability set associated with $\mathbb{M}$ is nonempty.

Collections of recursive conditional probabilities often arise in causal modelling – in particular, they are the foundation of the structural equation modelling approach Richardson and Robins (2013); Pearl (2009).

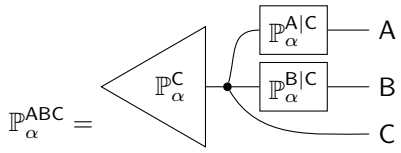
Note that validity is not a necessary condition for a conditional to define a non-empty probability set. The intuition for this is: if we have some $\mathbb{K} : X \twoheadrightarrow Y$, $\mathbb{K}$ might be an invalid $\mathsf{Y}|\mathsf{X}$ conditional on all of X, but might be valid on

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 6.14).

3.6.1 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 6.4.

Definition 3.28 (Conditional independence)**.** For a *probability model* $\mathbb{P}_\alpha$ and variables $\mathsf{A}, \mathsf{B}, \mathsf{Z}$, we say B is conditionally independent of A given C, written $\mathsf{B} \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{A}|\mathsf{C}$, if



(38)

Cho and Jacobs (2019) have shown that this definition coincides with the standard notion of conditional independence for a particular probability model (Theorem 6.15).

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

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

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

Definition 3.29 (Uniform conditional independence)**.** For a *probability set* $\mathbb{P}_{\{\}}$ and variables $\mathsf{A}, \mathsf{B}, \mathsf{C}$, we say B is uniformly conditionally independent of A given C, written $\mathsf{B} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{A}|\mathsf{C}$, if for all $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$ $\mathsf{B} \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{A}|\mathsf{C}$.

It is straightforward to show that uniform conditional independence satisfies the semi-graphoid axioms by application of Lemma 3.30

Lemma 3.30. $[\forall x : (f(x) \implies g(x))] \implies [(\forall x : f(x)) \implies (\forall x : g(x))]$

Proof. See appendix □

Lemma 3.31. *Given a standard measurable space $(\Omega, \mathcal{F})$ and $\mathbb{P}_{\{\}}$ on Ω, uniform conditional independence with respect to $\mathbb{P}_{\{\}}$ satisfies the semi-graphoid axioms.*

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

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

Then, by Lemma 3.30, $(\forall \mathbb{P}_\alpha \in \mathbb{P}_{\{\}} : f(\mathbb{P}_\alpha)) \implies (\forall \mathbb{P}_\alpha \in \mathbb{P}_{\{\}} : g(\mathbb{P}_\alpha))$.

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

4 Decision problems

We want to construct models to help make decisions. For our purposes, "making a decision" means we have some mathematically well-defined set C of choices under consideration, and some means of comparing one choice to another that induces a partial order on choices. If we are trying to be precise with language, we might call this *formal* decision making; people may often make decisions where it is difficult to identify a set C of choices under consideration or a rule for inducing a partial order on it.

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

We make the following assumptions about how choices are compared:

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

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

- There is a probability set $\mathbb{P}_\square$ representing "properties that hold regardless of which choice is taken"

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

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

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

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

Definition 4.1 (Probability gap model)**.** Given $(\Omega, \mathcal{F})$, a probability gap model is a triple $(\mathbb{P}_\square, \{\mathbb{P}_{\tilde{\alpha}}\}_A, f)$ where $\mathbb{P}_\square$ is a probability set and $\{\mathbb{P}_{\tilde{\alpha}}\}_A$ is a collection of probability sets and $f : A \to \mathcal{P}(\Delta(\Omega))$ is the map

$$\mathbb{P}_\alpha := f(\alpha) \tag{39}$$

$$= \mathbb{P}_\square \cap \mathbb{P}_{\tilde{\alpha}} \tag{40}$$

4.1 Conditional probability models

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

Definition 4.2 (Conditional probability model)**.** Given $(\Omega, \mathcal{F})$, $\mathsf{Y} : \Omega \to Y$, $\mathsf{X} : \Omega \to X$, a $\mathsf{Y}|\mathsf{X}$ conditional probability model is a probability gap model $(\mathbb{P}_\square^{\mathsf{Y}|\mathsf{X}[M]}, \{\mathbb{P}_\alpha^{\mathsf{X}[M]}\}_A, f)$.

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

Example 4.3 (Choice variable)**.** Suppose we have a procedure $\mathcal{C}$ that compares the elements of a countable set C and produces a choice according to some notion of the desirability of each one. Another procedure $\mathcal{Y}$ measures outcomes of interest. These are modelled with variables $\mathsf{C} : \Omega \to C$ and $\mathsf{Y} : \Omega \to Y$ respectively. Without loss of generality, we take $\Omega = C \times Y$ and C and Y are projections onto the matching sets.

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

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

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

$$\mathbb{P}_\alpha^{\mathsf{CY}}(B \times D) = \delta_\alpha(B)\mathbb{P}_\alpha^{\mathsf{Y}}(D) \tag{41}$$

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

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

Then

$$\mathbb{P}_\alpha^{\mathsf{CY}} = \mathbb{P}_\alpha^{\mathsf{C}} \odot \mathbb{P}_\square^{\mathsf{Y}|\mathsf{C}} \tag{42}$$

and so by Lemma 6.10

$$f(\alpha) = \mathbb{P}_\alpha \cap \mathbb{P}_\square \tag{43}$$

find a home for the following remarks

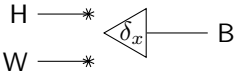
If the conditional probability $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ and all the marginal probabilities $\mathbb{P}_\alpha^{\mathsf{X}}$ are valid, then by Lemma 6.12 $\mathbb{P}_{\{\}} \cap \alpha \neq \emptyset$ for all $\alpha \in A$. Thus validity of all the individual parts is enough to ensure compatibility.

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

4.2 Example: invalidity

Body mass index is defined as a person's weight divided by the square of their height. Suppose we have a measurement 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

$$\nu^{\mathsf{B}|\mathsf{WH}} = \tag{44}$$



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

$$(\mathsf{B}, \mathsf{W}, \mathsf{H}) \bowtie \{(x, w, h)\} = \{\omega | (\mathsf{W}, \mathsf{H})(\omega) = (w, h), \mathsf{B}(\omega) = \frac{w}{h^2}\} \tag{45}$$

$$= \emptyset \tag{46}$$

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 44. We conjecture that in many cases this condition may either be trivial or implicitly taken into account when constructing conditional probabilities. However, we think it is useful to make this condition explicit nonetheless. We note that the invalid conditional probability 44 would be used to evaluate the causal effect of body mass index in the causal diagram found in Shahar (2009), presuming the author used the term "causal effect" to depend somehow on the function $x \mapsto P(\cdot|do(\mathsf{B} = x))$ as is the usual convention when discussing causal Bayesian networks.

End find a home for the following remarks

4.3 Response conditionals

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

$$\mathbb{P}^{\mathsf{Y}}_\alpha = \mathbb{P}^{\mathsf{X}}_\alpha \mathbb{P}^{\mathsf{Y}}_\square \qquad \forall \alpha \in A \tag{47}$$

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

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

4.4 Response conditionals and potential outcomes

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

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

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

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

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

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

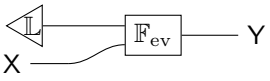
4.5 Randomness pushbacks

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

$$\mathbb{L}((y_i)_{i\in X}) = \prod_{i\in X} \mathbb{K}(y_i|i) \tag{48}$$

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

$$\mathbb{K} = \tag{49}$$



$$\iff \tag{50}$$

$$\mathbb{K}(A|x) = \int_{Y^X} \delta_{\mathrm{ev}(y^X,x)}(A)\mathbb{L}(\mathrm{d}y^X|x) \tag{51}$$

Unlike the case of function currying, $\mathbb{L}$ is not the unique Markov kernel for which 49 holds. In fact, we can substitute any $\mathbb{M}$ such that, for any $i \in X$

$$\sum_{y_{\{i\}^C} \in Y^{|X|-1}} \mathbb{M}((y_i)_{i \in X}) = \mathbb{K}(y_i|i) \tag{52}$$

This representation of a Markov kernel is called a *randomness pushback* by Fritz (2020). The idea is that the randomness in the orignal Markov kernel $\mathbb{L}$ has been "pushed back" to $\mathbb{L}$, which now passes through the deterministic Markov kernel $\mathbb{F}_{\text{ev}}$.

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

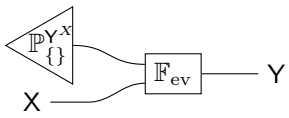
$$\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}\mathsf{Y}^X} = \mathbb{F}_{\text{ev}} \tag{53}$$

More commonly, this property is expressed as

$$\mathsf{Y} \overset{a.s.}{=} \text{ev}(\mathsf{X}, \mathsf{Y}^X) \tag{54}$$

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

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

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

(55)

Equation 55 is clearly a version of 49. As we have established, provided $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}$ exists, we can always introduce some variable Y^X and corresponding $\mathbb{P}_{\{\}}^{\mathsf{Y}^X}$ such that Equation 55 holds.

4.5.1 Choices aren't always known

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

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

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

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

4.6 Other decision theoretic causal models

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

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

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

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

The key difference is that Rohde and Lattimore's model employs different variables O and C to represent observations and interventional consequences,

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

If we require that the map from regimes to probability distributions is measurable (which can be trivially satisfied if the set of regimes is countable), then we can also characterise Dawid's approach using a probability set $\mathbb{P}_{\{\}}^{\mathsf{TY}|\mathsf{F}_T[M]}$ where F_T represents the prevailing regime, and T and Y are as in the quote above. Note that we do not consider conditional probability models here, because it is not obvious how we should say F_T is distributed given any choice of intervention; for any choice of intervention, F_T will sometimes take th evalue "observational" and sometimes take a value corresponding to the chosen intervention. Moreover, it seems inappropriate to posit a sequence of independent and identically distributed copies of F_T because we are likely to know it advance exactly when it will and won't take on the observational value, and so we can't associate it with a unique marginal distribution.

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

5 When do response conditionals exist?

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

We examine this question from two points of view: firstly, we aks *what kind of conditional probability models exhibit repeatable response condtionals?* Secondly, we ask *what kinds of experiments are these conditional probability*

models appropriate for? In the course of answering the second question, we show that apparently subtle differences in the description of an experimental procedure can determine whether a particular experiment should or should not be modeled with repeatable response conditionals.

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

5.1 Repeatable experiments

Our setup is a conditional probability model $(\mathbb{P}_{\square}^{\overline{\mathsf{Y}|\mathsf{D}}}, A)$ where $\mathsf{Y} := \mathsf{Y}_M = (\mathsf{Y}_i)_{i\in M}$ and $\mathsf{D} := \mathsf{D}_M = (\mathsf{D}_i)_{i\in M}$ for some index set M; we say such a model is a model of a *sequential experiment*. We say that Y_i is the consequence corresponding to the decision D_i for all $i \in M$ (i.e. variables with matching indices correspond). We say that $\mathbb{P}_{\square}^{\overline{\mathsf{Y}|\mathsf{D}}}$ features *repeatable response conditionals* if there exists a hypothesis H such that $\mathbb{P}_{\square}^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i} = \mathbb{P}_{\square}^{\mathsf{Y}_j|\mathsf{H}\mathsf{D}_j}$ for all $i, j \in M$, $\mathsf{H} \perp\!\!\!\perp_{\mathbb{P}_{\square}} \mathsf{D}$ and $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_{\square}} \mathsf{Y}_{M\setminus\{i\}}\mathsf{D}_{M\setminus\{i\}}|\mathsf{H}\mathsf{D}_i$. We remind the reader that in a conditional probability model, arbitrary conditional probabilities do not always exist, see Definition 3.17.

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

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

As we have mentioned, exchange commutativity is similar to the condition of *post-treatment exchangeability* found in Dawid (2020). Exchange commutativity is also very similar to the exchangeability assumption of GREENLAND and ROBINS (1986), and the assumption of exchangeability found in Banerjee et al.

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

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

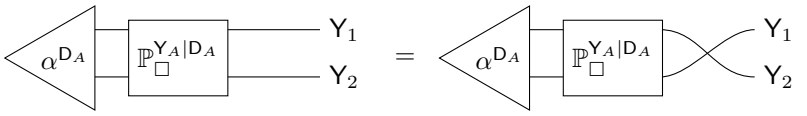
"("SUTVA) comprises two subassumptions. First, it assumes that *there is no interference between units (Cox 1958)*; that is, neither $Y_i(1)$ nor $Y_i(0)$ is affected by what action any other unit received. Second, it assumes that *there are no hidden versions of treatments*; no matter how unit i received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

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

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

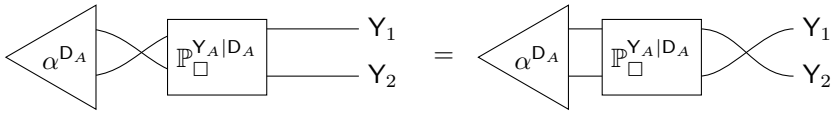
The difference is illustrated by the following pair of diagrams.

Exchangeability (swapping labels):



(56)

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



(57)

—-end not sure where to put——

5.2 Consequence contractibility

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

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

Definition 5.1 (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 5.2 (Exchange commutativity)**.** Suppose we have a sample space $(\Omega, \mathcal{F})$ and a conditional probability model $(\mathbb{P}_\square^{\mathsf{Y}|\mathsf{D}[M]}, \{\mathbb{P}_{\tilde{\alpha}}^{\mathsf{D}}\}_A, f)$ with $\mathsf{Y} := \mathsf{Y}_M := (\mathsf{Y}_i)_M$, $\mathsf{D} := \mathsf{D}_M := (\mathsf{D}_i)_M$, $M \subseteq \mathbb{N}$. If, for any decision rule $\alpha \in A$,

$$\mathbb{P}_{\tilde{\alpha}}^{\mathsf{D}} \odot \mathrm{swap}_{\rho(\mathsf{D})} \mathbb{P}_\square^{\mathsf{Y}|\mathsf{D}} = \mathbb{P}_{\tilde{\alpha}}^{\mathsf{D}} \odot \mathbb{P}_\square^{\mathsf{Y}|\mathsf{D}} \mathrm{swap}_{\rho(\mathsf{Y})} \tag{58}$$

Then $\mathbb{P}_\square$ *commutes with exchange* with respect to (D, Y).

A conditional probability model exhibits locality of consequences if, given two different choices that agree on an subsequence of indices, the model yields identical outcomes if we restrict our attention to the subsequence on which the different choices match. For example, if we have $\mathsf{D} = (\mathsf{D}_1, \mathsf{D}_2, \mathsf{D}_3)$ and $\mathsf{Y} = (\mathsf{Y}_1, \mathsf{Y}_2, \mathsf{Y}_3)$ and $\mathbb{P}_{\tilde{\alpha}}^{\mathsf{D}_1\mathsf{D}_3} = \mathbb{P}_{\tilde{\beta}}^{\mathsf{D}_1\mathsf{D}_3}$ then $\mathbb{P}_{\alpha}^{\mathsf{Y}_1\mathsf{D}_1\mathsf{Y}_3\mathsf{D}_3} = \mathbb{P}_{\beta}^{\mathsf{Y}_1\mathsf{D}_1\mathsf{Y}_3\mathsf{D}_3}$.

Definition 5.3 (Locality of consequences)**.** Suppose we have a sample space $(\Omega, \mathcal{F})$ and a conditional probability model $(\mathbb{P}_\square^{\mathsf{Y}|\mathsf{D}[M]}, \{\mathbb{P}_{\tilde{\alpha}}^{\mathsf{D}}\}_A, f)$ with $\mathsf{Y} := \mathsf{Y}_M := (\mathsf{Y}_i)_M$, $\mathsf{D} := \mathsf{D}_M := (\mathsf{D}_i)_M$, $M \subseteq \mathbb{N}$. For any ordered sequence $S = (s_i)_{i \in Q}$ where $Q \subset M$ and $i < j \implies s_i < s_j$, let $\mathsf{D}_S := (\mathsf{D}_i)_{i \in S}$ and $\mathsf{D}_T := (\mathsf{D}_i)_{i \in T}$. If for any $\alpha, \beta \in R$

$$\mathbb{P}_{\alpha}^{\mathsf{D}_S} = \mathbb{P}_{\beta}^{\mathsf{D}_S} \tag{59}$$

$$\implies \mathbb{P}_{\alpha}^{(\mathsf{D}_i, \mathsf{Y}_i)_{i \in S}} = \mathbb{P}_{\beta}^{(\mathsf{D}_i, \mathsf{Y}_i)_{i \in S}} \tag{60}$$

then $\mathbb{P}_\square$ exhibits *locality of consequences* with respect to (D, Y).

Neither condition implies the other.

Lemma 5.4. *Exchange commutativity does not imply locality of consequences or vise versa.*

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

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

$$\mathbb{P}_{\square}^{\mathsf{Y}_1\mathsf{Y}_2|\mathsf{D}_1\mathsf{D}_2}(y_1, y_2|d_1, d_2) = [\![(y_1, y_2) = (d_1 + d_2, d_1 + d_2)]\!] \tag{61}$$

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 $\mathbb{P}_{\square}$ exhibits non-local consequences. However, taking $(d_i, d_j) := \delta_{d_i d_j} \in A$,

$$\begin{aligned}
\mathbb{P}_{d_2,d_1}^{\mathsf{Y}_1\mathsf{D}_1\mathsf{Y}_2\mathsf{D}_2}(y_1, d_1, y_2, d_2) &= [\![(y_1, y_2) = (d_2 + d_1, d_2 + d_1)]\!] &(62)\\
&= [\![(y_2, y_1) = (d_1 + d_2, d_1 + d_2)]\!] &(63)\\
&= \mathbb{P}_{d_1,d_2}^{\mathsf{Y}_1\mathsf{D}_1\mathsf{Y}_2\mathsf{D}_2}(y_2, d_2, y_1, d_1) &(64)
\end{aligned}$$

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

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

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

$$\mathbb{P}_{\square}\mathsf{Y}_1\mathsf{Y}_2|\mathsf{D}_1\mathsf{D}_2(y_1, y_2|d_1, d_2) = [\![(y_1, y_2) = (0, 1)]\!] \tag{65}$$

for all $\alpha \in A$.

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

$$\begin{aligned}
\mathbb{P}_{\alpha}^{\mathsf{Y}_S\mathsf{D}_S}(y_s, d_s) &= \sum_{y_2' \in \{0,1\}^{S^C}} [\![(y_1, y_2) = (0, 1)]\!]\mathbb{P}_{\alpha}^{\mathsf{D}_S}(d_s) &(66)\\
&= \mathbb{P}_{\beta}^{\mathsf{Y}_S\mathsf{D}_S}(y_s, d_s) &(67)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}_{\alpha}\mathsf{Y}_1\mathsf{Y}_2(y_1, y_2) &= [\![(y_1, y_2) = (0, 1)]\!] &(68)\\
&\neq \mathbb{P}_{\beta}\mathsf{Y}_1\mathsf{Y}_2(y_2, y_1) &(69)
\end{aligned}$$

□

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

- The marginal distribution of both experiments in isolation is Bernoulli(0.5) no matter what choices I make, so a model of this experiment would satisfies Definition 5.3
- Nevertheless, the choice for the first experiment affects the result of the second experiment

Note that this example would not satisfy exchange commutativity.

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

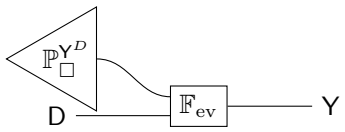
Definition 5.5 (Consequence contractibility)**.** A conditional probability model $(\mathbb{P}_\square^{\overline{\mathsf{Y}|\mathsf{D}}}, A)$ is causally contractible if it is both commutative with exchange and commutative with marginalisation.

5.3 Repeatable response conditionals exist iff a model is consequence contractible

The main result in this section is Theorem 5.7 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 $\mathbb{P}_\square^{\mathsf{Y}_1|\mathsf{D}_1\mathsf{H}}$. Note the hypothesis H that appears in this conditional; it can be given the interpretation of a random variable that expresses the "true but initially unknown" $\mathsf{Y}_1|\mathsf{D}_1$ conditional probability.

Lemma 5.6 (Exchangeable randomness pushback)**.** *Given a conditional probability model $(\mathbb{P}_\square^{\mathsf{Y}|\mathsf{D}}, A)$ such that $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$, $\mathbb{P}_\square$ is causally contractible if and only if there exists a column exchangeable probability distribution $\mathbb{P}^{\mathsf{Y}^D}$ such that*

$$\mathbb{P}_\square^{\mathsf{Y}|\mathsf{D}} = \tag{70}$$



$$\iff \tag{71}$$

$$\mathbb{P}_\square^{\mathsf{Y}|\mathsf{D}}(y|d) = \mathbb{P}^{(\mathsf{Y}^D_{d_i i})_{\mathbb{N}}}(y) \tag{72}$$

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

$$ev : D^{\mathbb{N}} \times Y^{D\times\mathbb{N}} \to Y \tag{73}$$

$$((d_i)_{\mathbb{N}}, (y_{ij})_{i\in D, j\in\mathbb{N}}) \mapsto (y_{d_i i})_{i\in\mathbb{N}} \tag{74}$$

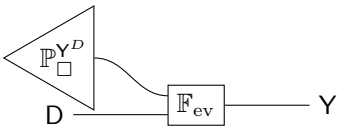
Proof. Only if: Choose $e := (e_i)_{i\in\mathbb{N}}$ such that $e_{|D|i+j}$ is the ith element of D for all $i, j \in \mathbb{N}$. Abusing notation, write e also for the decision function that chooses e deterministically.

Define

$$\mathbb{P}^{\mathsf{Y}^D}((y_{ij})_{D\times\mathbb{N}}) := \mathbb{P}_e^{\mathsf{Y}}((y_{|D|i+j})_{i\in D, j\in\mathbb{N}}) \tag{75}$$

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

$$\mathbb{Q} : D \nrightarrow Y \tag{76}$$

$\mathbb{Q} :=$ 

(77)

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

$$\sum_{y\in\mathsf{Y}^{-1}(y_A)} \mathbb{Q}(y|d) = \sum_{y\in\mathsf{Y}^{-1}(y_A)} \mathbb{P}^{(\mathsf{Y}^D_{d_i i})_A}(y) \tag{78}$$

$$= \sum_{y\in\mathsf{Y}^{-1}(y_A)} \mathbb{P}_e^{(\mathsf{Y}_{|D|d_i+i})_A}(y) \tag{79}$$

$$= \mathbb{P}_e^{\mathsf{Y}_B}(y_A) \tag{80}$$

$$= \mathbb{P}_d^{\mathsf{Y}_A}(y_A) \qquad \text{by causal contractibility} \tag{81}$$

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

$$\mathbb{Q}(y|d) = \mathbb{P}_d^{\mathsf{Y}}(y) \tag{82}$$

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

$$\mathbb{Q}(y|d) = \mathbb{P}_d^{\mathsf{Y}|\mathsf{D}}(y|d) \tag{83}$$

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

$$\mathbb{P}^{\mathsf{Y}|\mathsf{D}} = \mathbb{Q} \tag{84}$$

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

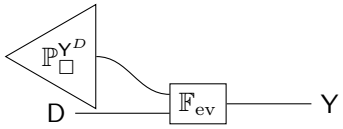
$$\mathbb{P}^{\mathsf{Y}_S^D} = \mathbb{P}_e^{\mathsf{Y}_{\rho(S)}}) \tag{85}$$

$$= \mathbb{P}_e^{\mathsf{Y}_{\rho(T)}}) \quad \text{by contractibility of } \mathbb{P} \text{ and the equality } e_{\rho(S)} = e_{\rho(T)} \tag{86}$$

$$= \mathbb{P}^{\mathsf{Y}_T^D} \tag{87}$$

If: Suppose

$$\mathbb{P}^{\mathsf{Y}|\mathsf{D}} = \tag{88}$$



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

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

By definition,

$$\mathbb{P}^{\mathsf{Y}_S|\mathsf{D}}(y_S|d) = \sum_{y_S^D \in Y^{|D| \times |S|}} \mathbb{P}^{\mathsf{Y}_S^D}(y_S^D) \mathbb{F}_{\text{ev}}(y_S|d, y_S^D) \tag{89}$$

$$= \sum_{y_S^D \in Y^{|D| \times |T|}} \mathbb{P}^{\mathsf{Y}_T^D}(y_S^D) \mathbb{F}_{\text{ev}}(y_S|d, y_S^D) \quad \text{by contractibility of } \mathbb{P}^{\mathsf{Y}_T^D} \tag{90}$$

$$= \mathbb{P}^{\mathsf{Y}_T|\mathsf{D}}(y_S|d) \tag{91}$$

□

As we pointed out, there are similarities between tabular distributions like $\mathbb{P}^{\mathsf{Y}^D}$ that appears in Lemma 5.6 and potential outcomes causal models. However, the $\mathbb{P}^{\mathsf{Y}^D}$ that appears in this lemma usually can't be interpreted as a distribution of potential outcomes. For example, consider a series of bets on fair coinflips. Model the consequence Y_i as uniform on $\{0, 1\}$ for any decision D_i, for all i. Specifically, $D = Y = \{0, 1\}$ and $\mathbb{P}_\alpha^{\mathsf{Y}_n}(y) = \prod_{i \in [n]} 0.5$ for all n, $y \in Y^n$, $\alpha \in R$. Then the construction of $\mathbb{P}^{\mathsf{Y}^D}$ following the method in Lemma 5.6 yields $\mathbb{P}^{Y_i^D}(y_i^D) = \prod_{j \in D} 0.5$ for all $y_i^D \in Y^D$. In this model Y_i^0 and Y_i^1 are independent and uniformly distributed. However, if we wanted Y_i^0 to be interpretable as "what would happen if I bet on outcome 0 on turn i" and Y^1 to represent "what would happen if I bet on outcome 1 on turn i", then we ought to have $\mathsf{Y}_i^0 = 1 - \mathsf{Y}_i^1$.

Lemma 5.6 also does not establish that causal contractibility is necessary for the existence of a potential outcomes. A counterexample is any potential outcomes model with potential outcomes Z^D where the distribution $\mathbb{P}^{\mathsf{Z}^D}$ is not column exchangeable. Such a model is not causally contractible.

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

Theorem 5.7. *Suppose we have a sample space $(\Omega, \mathcal{F})$ and a conditional probability model $(\mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{D}}, A)$ such that $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$. $\mathbb{P}_{\square}$ is causally contractible if and only if there exists some $\mathsf{H} : \Omega \to H$ such that $\mathbb{P}_{\square}^{\mathsf{H}}$ and $\mathbb{P}_{\square}^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i}$ exist for all $i \in \mathbb{N}$ and*

$$\mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{D}} = \quad \mathsf{H},\ \mathsf{D} \longrightarrow \boxed{\Pi_{D,i} \longrightarrow \boxed{\mathbb{P}_{\square}^{\mathsf{Y}_0|\mathsf{H}\mathsf{D}_0}} \longrightarrow \mathsf{Y}_i \quad i \in \mathbb{N}} \tag{92}$$

$$\iff \tag{93}$$

$$\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_{\square}} \mathsf{Y}_{\mathbb{N}\setminus i}, \mathsf{D}_{\mathbb{N}\setminus i} | \mathsf{H}\mathsf{D}_i \qquad \forall i \in \mathbb{N} \tag{94}$$

$$\wedge \mathsf{H} \perp\!\!\!\perp_{\mathbb{P}_{\square}} \mathsf{D} \tag{95}$$

$$\wedge \mathbb{P}_{\square}^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i} = \mathbb{P}^{\mathsf{Y}_0|\mathsf{H}\mathsf{D}_0} \qquad \forall i \in \mathbb{N} \tag{96}$$

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

Proof. We make use of Lemma 5.6 to show that we can represent the conditional probability $\mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{D}}$ as

$$\mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{D}} = \quad \boxed{\mathbb{P}_{\square}^{\mathsf{Y}^D}},\ \mathsf{D} \longrightarrow \boxed{\mathbb{F}_{\text{ev}}} \longrightarrow \mathsf{Y} \tag{97}$$

$$\tag{98}$$

As a preliminary, we will show

$$\mathbb{F}_{\text{ev}} = \quad \mathsf{H},\ \mathsf{D} \longrightarrow \boxed{\Pi_{Y^D,i},\ \Pi_{D,i} \longrightarrow \boxed{\mathbb{F}_{\text{ev},i}} \longrightarrow \mathsf{Y}_i \quad i \in \mathbb{N}} \tag{99}$$

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

$$((y_i)_{i\in D}, d) \mapsto y_d \tag{100}$$

Recall that ev is the function

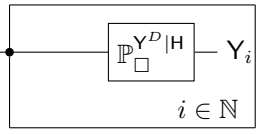
$$((d_i)_{\mathbb{N}}, (y_{ij})_{i\in D, j\in\mathbb{N}}) \mapsto (y_{d_i i})_{i\in\mathbb{N}} \tag{101}$$

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

$$\begin{aligned}
\mathbb{F}_{\mathrm{ev}}(\prod_{i\in\mathbb{N}} A_i | (d_i)_{\mathbb{N}}, (y_{ij})_{i\in D, j\in\mathbb{N}}) &= \delta_{(y_{d_i i})_{i\in\mathbb{N}}}(\prod_{i\in\mathbb{N}} A_i) && (102)\\
&= \prod_{i\in\mathbb{N}} \delta_{y_{d_i i}}(A_i) && (103)\\
&= \mathrm{copy}^{\mathbb{N}} \prod_{i\in\mathbb{N}} (\mathit{\Pi}_{D,i} \otimes \mathit{\Pi}_{Y,i}) \mathbb{F}_{\mathrm{ev}_{Y^D \times D}} && (104)
\end{aligned}$$

Which is what we wanted to show.

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

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

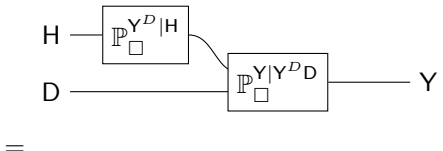
(105)

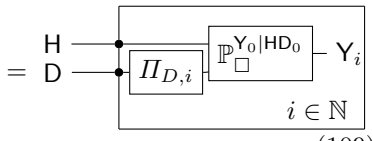
$$\iff \tag{106}$$

$$\mathbb{P}_{\square}^{\mathsf{Y}^D|\mathsf{H}}(\prod_{i\in\mathbb{N}} A_i | h) = \prod_{i\in\mathbb{N}} \mathbb{P}_{\square}^{\mathsf{Y}_0^D|\mathsf{H}}(A_i | h) \tag{107}$$

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

$$\mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{HD}} = \mathbb{P}_{\square}^{\mathsf{Y}^D|\mathsf{HD}} \odot \mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{Y}^D\mathsf{HD}} \tag{108}$$



$=$ 

(109)

If: By assumption

$$\mathbb{P}_{\square}^{\mathsf{Y}|\mathsf{D}}(\prod_{i\in\mathbb{N}} A_i | h, (d_i)_{i\in\mathbb{N}}) = \int_H \prod_{i\in\mathbb{N}} \mathbb{P}_{\square}^{\mathsf{Y}_1|\mathsf{HD}_1}(A_i | h, d_i) \mathbb{P}_{\square}^{\mathsf{H}}(\mathrm{d}h) \tag{110}$$

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

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{Y}_M}(A) &= \int_{D^{\mathbb{N}}} \mathbb{P}_\alpha^{\mathsf{Y}_M|\mathsf{D}}(A|d)\mathbb{P}_\alpha^{\mathsf{D}}(\mathrm{d}d) && (111)\\
&= \int_H \int_{D^{\mathbb{N}}} \prod_{i\in M} \mathbb{P}_\square^{\mathsf{Y}_1|\mathsf{HD}_1}(A_i|h,d_i)\mathbb{P}_\alpha^{\mathsf{D}}(\mathrm{d}d)\mathbb{P}_\square^{\mathsf{H}}(\mathrm{d}h) && (112)\\
&= \int_H \int_{D^{|M|}} \prod_{i\in M} \mathbb{P}_\square^{\mathsf{Y}_1|\mathsf{HD}_1}(A_i|h,d_i)\mathbb{P}_\alpha^{\mathsf{D}_M}(\mathrm{d}d_M)\mathbb{P}_\square^{\mathsf{H}}(\mathrm{d}h) && (113)\\
&= \int_H \int_{D^{|M|}} \prod_{i\in M} \mathbb{P}_\square^{\mathsf{Y}_1|\mathsf{HD}_1}(A_i|h,d_i)\mathbb{P}_{\alpha'}^{\mathsf{D}_N}(\mathrm{d}d_N)\mathbb{P}_\square^{\mathsf{H}}(\mathrm{d}h) && (114)\\
&= \int_H \int_{D^{\mathbb{N}}} \prod_{i\in M} \mathbb{P}_\square^{\mathsf{Y}_1|\mathsf{HD}_1}(A_i|h,d_i)\mathbb{P}_{\alpha'}^{\mathsf{D}}(\mathrm{d}d)\mathbb{P}_\square^{\mathsf{H}}(\mathrm{d}h) && (115)\\
&= \mathbb{P}_{\alpha'}^{\mathsf{Y}_M}(A) && (116)
\end{aligned}$$

□

5.4 Modelling different measurement procedures

I've started but not finished revising the following

When is it reasonable to assume causal contractibility? This involves two judgements – that choices and outcomes can be paired up so that each choice only influences the corresponding outcome, and that there is an equivalence between exchanging choices and exchanging outcomes. Here, we're going to focus on when it is reasonable to assume exchange commutativity.

There is a superficially plausible but unsound 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 exchange commutativity holds. The problem with this line of reasoning is that we cannot deduce from the interchangeability of "experimental units" that exchanging choices and exchanging outcomes are equivalent. For example, if choices depend on some feature of the "experimental units", then exchanging choices will lead to choices having different dependence on experimental units.

This is exacerbated by the fact that exchange of experimental units is an operation on a *measurement procedure*. Measurement procedures involve interaction with the real world and we do not have a method for precisely describing them (unlike probablistic models). Thus incomplete descriptions of a measurement procedure often leave open the possibility that choices do depend on properties of experimental units. We can sometimes rule this possibility out by specifying a measurement procedure that only allows for *deterministic* choices.

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

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

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

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

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

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

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

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

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

Consider the following two scenarios:

1. Dr Alice is going to see two patients who are both complaining of lower back pain and are otherwise unknown to Alice. Prior to seeing them, she settles on a decision function α which deterministically sets her treatment choices according to a function decisions(α)

2. As before, but α is a "decision inclination" and $\mathbb{P}_{\alpha}^{\mathsf{D}_1\mathsf{D}_2}$ nondeterministic

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

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

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

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

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

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

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

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

Make the assumption that, on the basis that the patients are indistinguishable to Alice at the time of model construction, the same model is appropriate

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

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

I've started but not finished revising the previous

5.6 Causal consequences of non-deterministic variables

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

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

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

The determinism assumption rules out choosing decisions randomly. However, if we have response conditionals with a particular conditioning variable, response conditionals for other conditioning variables may exist if a certain conditional independence that we refer to as *proxy control* holds. That is, if we have a response conditional for (X, Y) given D, D is deterministic for all choices and Y is independent of D given X, then we also have a response conditional for Y given X and X may not be deterministic.

We also show that proxy control is necessary for the existence of additional response conditionals if D is deterministically controllable; that is, if it can be forced to take on any deterministic probability distribution. If the judgements underpinning the existence of response conditionals ultimately rest on decision variables that are deterministic for each choice that can be made, and we claim that a response conditional for Y given X exists where X is just some not-necessarily-deterministic variable, then X must be a proxy for controlling Y given D.

Definition 5.8 (Deterministically controllable)**.** Given a probabilty gap model $(\mathbb{P}_\square, \{\mathbb{P}^{\mathsf{D}}_\alpha\}_A, f)$ on $(\Omega, \mathcal{F})$ and a variable $\mathsf{X} : \Omega \to X$, if for any $x \in X$ there exists $\alpha \in A$ such that $\mathbb{P}^{\mathsf{X}}_\alpha = \delta_x$ then X is determinstically controllable.

Theorem 5.9. *Given* $(\mathbb{P}^{\mathsf{XY}|\mathsf{D}[M]}_\square, \{\mathbb{P}^{\mathsf{D}}_\alpha\}_A, f)$ *with decisions* D_M *and consequences* $\mathsf{Y}_M, \mathsf{X}_M$*, if* $\mathbb{P}^{\mathsf{Y}_M\mathsf{X}_M|\mathsf{D}_M}_\square$ *is causally contractible with response conditional* $\mathbb{P}^{\mathsf{Y}_0\mathsf{X}_0|\mathsf{D}_0\mathsf{H}}_\square$ *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* $\mathbb{P}^{\mathsf{Y}_M|\mathsf{X}_M}_\square$ *exists. If* D_M *is deterministically controllable and* D *countable, then* $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_\square} \mathsf{D}_i|\mathsf{H}\mathsf{X}_i$ *is also necessary for the existence of* $\mathbb{P}^{\mathsf{Y}_M|\mathsf{X}_M}_\square$*.*

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

From causal contractibility we have

$$(\mathsf{X}_i, \mathsf{Y}_i) \perp\!\!\!\perp_{\mathbb{P}_\square} (\mathsf{X}_{\{i\}^C}, \mathsf{Y}_{\{i\}^C}, \mathsf{D}_{\{i\}^C}) | \mathsf{H}\mathsf{D}_i \tag{117}$$

$$\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_\square} (\mathsf{Y}_{\{i\}^C}, \mathsf{X}_{\{i\}^C}) | \mathsf{H}\mathsf{D}_i\mathsf{X}_i \tag{118}$$

Where Eq. 118 follows from 117 by weak union.

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 6.22 and the existence of $\mathbb{P}^{\mathsf{Y}_i\mathsf{X}_i|\mathsf{H}\mathsf{D}_i}$ for all $i \in M$, $\mathbb{P}_\square^{\mathsf{Y}_i|\mathsf{H}\mathsf{X}_i}$ exists for all i. Furthermore, because $\mathbb{P}^{\mathsf{Y}_i\mathsf{X}_i|\mathsf{H}\mathsf{D}_i} = \mathbb{P}^{\mathsf{Y}_j\mathsf{X}_j|\mathsf{H}\mathsf{D}_j}$ for all $i, j \in M$, $\mathbb{P}_\square^{\mathsf{Y}_i|\mathsf{H}\mathsf{X}_i} = \mathbb{P}_\square^{\mathsf{Y}_j|\mathsf{H}\mathsf{X}_j}$ for all $i, j \in M$. **Necessity:** We will show for all $\alpha \in A$, $B \in \mathcal{Y}$, $(x, d, h) \in X \times D \times H$ that

$$\mathbb{P}_\alpha^{\mathsf{Y}_0|\mathsf{X}_0\mathsf{D}_0\mathsf{H}}(B|x, d, h) = \mathbb{P}_\alpha^{\mathsf{Y}_0|\mathsf{X}_0\mathsf{H}}(B|x, h) \tag{119}$$

By assumption, we have the conditionals $\mathbb{P}_\square^{\mathsf{Y}_i\mathsf{X}_i|\mathsf{D}_i\mathsf{H}}$ and $\mathbb{P}_\square^{\mathsf{X}_i|\mathsf{H}\mathsf{D}_i}$ for all $i \in M$. We can conclude that $\mathbb{P}_\square^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{D}_0 i\mathsf{H}}$ also exists, as it is a higher order conditional with respect to $\mathbb{P}_\square^{\mathsf{Y}_i\mathsf{X}_i|\mathsf{D}_i\mathsf{H}}$.

For arbitrary $d \in D$, let $\alpha_d \in A$ be such that $\mathbb{P}_{\alpha_d}^{\mathsf{D}_i} = \delta_d$. For every version of $\mathbb{P}_{\alpha_d}^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{D}_i\mathsf{H}}$ and $\mathbb{P}_{\alpha_d}^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{D}_i\mathsf{H}}$

$$\mathbb{P}_{\alpha_d}^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{H}}(B|x, h) = \int_D \mathbb{P}_{\alpha_d}^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{D}_i\mathsf{H}}(B|x, d', h)\delta_d(\mathrm{d}d') \tag{120}$$

$$= \mathbb{P}_{\alpha_d}^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{D}_i\mathsf{H}}(B|x, d, h) \tag{121}$$

For all $x \in X$, $h \in H$ $B \subset \mathcal{Y}$ except on a set of points $C \subset X \times H$ of uniform $\mathbb{P}_{\alpha_d}$ measure 0.

Need to add independence of hypothesis to representation theorem

However, note that for any α

$$\mathbb{P}_\alpha^{\mathsf{X}_i\mathsf{H}\mathsf{D}_i}(E \times F \times G) = \sum_{d \in G} \mathbb{P}_\alpha^{\mathsf{D}_i}(d)\mathbb{P}_\square^{\mathsf{X}_i\mathsf{H}|\mathsf{D}_i}(E \times F|d) \tag{122}$$

$$= \sum_{d \in G} \mathbb{P}_\alpha^{\mathsf{D}_i}(d) \sum_{d' \in D} \mathbb{P}_\square^{\mathsf{X}_i\mathsf{H}|\mathsf{D}_i}(E \times F|d')\mathbb{P}_{\alpha_d}^{\mathsf{D}_i}(\{d'\} \tag{123}$$

$$= \sum_{d \in G} \mathbb{P}_\alpha^{\mathsf{D}_i}(d)\mathbb{P}_{\alpha_d}^{\mathsf{X}_i\mathsf{H}\mathsf{D}_i}(E \times F \times \{d\}) \tag{124}$$

Thus for each $d \in D$ the set $\{d\} \times C \subset D \times X \times H$ is of uniform $\mathbb{P}_\alpha$ measure 0 for any $\alpha \in A$. Because $\mathbb{P}_\square = \cup_{\alpha \in A}\mathbb{P}_\alpha$, it is also of uniform $\mathbb{P}_\square$ measure 0. Thus

$$\mathbb{P}_\square^{\mathsf{Y}_0|\mathsf{X}_0\mathsf{H}}(B|x, h) = \mathbb{P}_\square^{\mathsf{Y}_0|\mathsf{X}_0\mathsf{D}_0\mathsf{H}}(B|x, d, h) \tag{125}$$

as desired. □

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.

5.7 Body mass index revisited

We return briefly to consider the question: given some collection of people indexed by M, with body mass index B_i and health outcomes of interest Y_i and some choices D_i a decision maker is contemplating relevant to these characteristics, suppose we have a conditional probability model $(\mathbb{P}_{\square}^{\mathsf{BY}|\mathsf{D}[M]}, \{\mathbb{P}_{\tilde{\alpha}}^{\mathsf{D}}\}_A, f)$ causally contractible with respect to (D, Y) (for example, perhaps decision maker is contemplating a treatment plan to apply to every individual).

Do response conditionals $\mathbb{P}_{\square}^{\mathsf{Y}_i|\mathsf{B}_i}$ exist? We have by Lemma 5.9 that this exists if and only if $\mathsf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_{\square}} \mathsf{D}_i | \mathsf{B}_i$. Thus we have reduced the question of the existence of response conditionals for BMI (or "causal effects" of BMI) to an empirical question. We might guess this is unlikely to hold; not only are there multiple ways we could imagine affecting a person's BMI with possibly different health implications, but it seems unlikely that the ultimate health outcome someone experiences can be predicted from BMI alone.

However, there might be something to be said for a "causal effect of BMI". In particular, while it seems unlikely that BMI is a precise proxy for controlling health outcomes, it seems to at least be a reasonable empirical question to ask if BMI is an *approximate* proxy for health outcomes.

Do I prove a theorem about approximate proxy control?

6 Appendix, needs to be organised

6.1 Existence of conditional probabilities

Lemma 6.1 (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 3.11, for any $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$

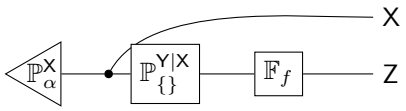
$$\mathbb{P}_\alpha^{\mathsf{XZ}} = \mathbb{P}_\alpha^{\mathsf{XY}} \mathbb{F}_{\mathrm{id}_X \otimes f} \tag{126}$$

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

$$\begin{aligned} \mathbb{F}_{\mathrm{id}_X \otimes f}(A \times B | x, y) &= \delta_x(A) \delta_{f(y)}(B) && (127)\\ &= \mathbb{F}_{\mathrm{id}_X}(A|x) \otimes \mathbb{F}_f(B|y) && (128)\\ \implies \mathbb{F}_{\mathrm{id}_X \otimes f} &= \mathbb{F}_{\mathrm{id}_X} \otimes \mathbb{F}_f && (129) \end{aligned}$$

Thus

$$\mathbb{P}_\alpha^{\mathsf{XZ}} = (\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}) \mathbb{F}_{\mathrm{id}_X} \otimes \mathbb{F}_f \tag{130}$$



$$= \tag{131}$$

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 6.2 (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 6.3 (Existence of higher order valid conditionals with respect to probability sets)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_{\{\}}$ with regular conditional $\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$ and Y and Z standard measurable. Then there exists a regular $\mathbb{P}_{\{\}}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$.*

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

$$n(f^{-1}(y)|x, y) = 1 \tag{132}$$

$$m(\mathsf{Y}^{-1}(A) \cap B | x) = \int_A n(B|x, y) m\mathbb{F}_f(dy|x) \forall A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z} \tag{133}$$

In particular, $\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$ is a Borel measurable map $X \nrightarrow Y \times Z$. Thus equation 133 implies for all $A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z}$

$$\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}(\mathsf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x,y)\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}\mathbb{F}_f(dy|x) \tag{134}$$

$$= \int_A n(B|x,y)\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}}(dy|x) \tag{135}$$

Where Equation 135 follows from Lemma 6.1.
Then, for any $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$

$$\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}(\mathsf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x,y)\mathbb{P}_{\alpha}^{\mathsf{Y}|\mathsf{X}}(dy|x) \tag{136}$$

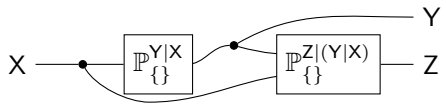
which implies n is a version of $\mathbb{P}_{\{\}}^{\mathsf{YZ}|(\mathsf{Y}|\mathsf{X})}$. By Lemma 6.1, $n\mathbb{F}_f$ is a version of $\mathbb{P}_{\{\}}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$. □

We might be motivated to ask whether the higher order conditionals in Theorem 6.3 can be chosen to be valid. Despite Lemma 6.8 showing that the existence of proper conditional probabilities implies the existence of valid ones, we cannot make use of this in the above theorem because Equation 132 makes n proper with respect to the "wrong" sample space $(Y \times Z, \mathcal{Y} \otimes \mathcal{Z})$ while what we would need is a proper conditional probability with respect to $(\Omega, \mathcal{F})$.

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

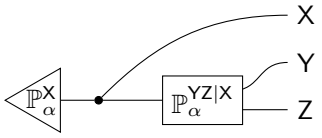
Theorem 6.4 (Higher order conditionals)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_{\{\}}$ with conditional $\mathbb{P}_{\{\}}^{\mathsf{YZ}|\mathsf{X}}$. Then $\mathbb{P}_{\{\}}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ is a version of $\mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{YX}}$*

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

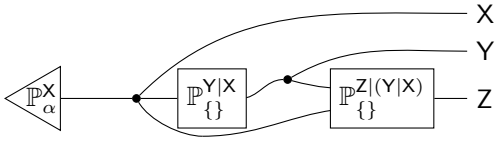


$$\mathbb{P}^{\mathsf{YZ}|\mathsf{X}}_{\alpha} = \tag{137}$$

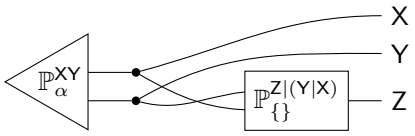
$$\implies \mathbb{P}^{\mathsf{XYZ}}_{\alpha} = \tag{138}$$



$$= \tag{139}$$



$$= \tag{140}$$

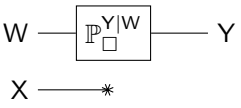


Thus $\mathbb{P}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}_{\{\}}$ is a version of $\mathbb{P}^{\mathsf{Z}|\mathsf{YX}}_{\alpha}$ for all α and hence also a version of $\mathbb{P}^{\mathsf{Z}|\mathsf{YX}}_{\{\}}$. □

Theorem 6.5. *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 6.15, for each $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$ there exists $\mathbb{P}^{\mathsf{Z}|\mathsf{Y}}_{\alpha}$ such that

$$\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}_{\alpha} = \tag{141}$$



□

Theorem 6.6 (Valid higher order conditionals)**.** *Suppose we have a sample space* $(\Omega, \mathcal{F})$*, variables* $\mathsf{X} : \Omega \to X$ *and* $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ *and a probability set* $\mathbb{P}_{\{\}}$ *with regular conditional* $\mathbb{P}^{\mathsf{YZ}|\mathsf{X}}_{\{\}}$*, Y discrete and Z standard measurable. Then there exists a valid regular* $\mathbb{P}^{\mathsf{Z}|\mathsf{XY}}_{\{\}}$.

Proof. By Theorem 6.3, we have a higher order conditional $\mathbb{P}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}_{\{\}}$ which, by Theorem 6.4 is also a version of $\mathbb{P}^{\mathsf{Z}|\mathsf{XY}}_{\{\}}$.

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

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

$$\forall A \in \mathcal{Z}, x \in X, y \in Y : \mathbb{P}^{\mathsf{YZ}|\mathsf{X}}_{\{\}}(\{y\} \times A|x) = 0 \implies (\mathbb{Q}(A|x, y) = 0) \vee ((\mathsf{X}, \mathsf{Y}) \bowtie \{(x, y)\} = \emptyset) \tag{142}$$

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

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

6.2 Validity

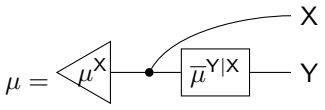
Validity is related to *proper* conditional probabilities. In particular, valid conditional probabilities exist when regular proper conditional probabilities exist.

Definition 6.7 (Regular proper conditional probability)**.** Given a probability space $(\mu, \Omega, \mathcal{F})$ and a variable $\mathsf{X} : \Omega \to X$, a regular proper conditional probbability $\mu^{|\mathsf{X}} : X \to\!\!\!\!\bullet\ \Omega$ is Markov kernel such that

$$\mu(A \cap \mathsf{X}^{-1}(B)) = \int_B \mu^{|\mathsf{X}}(A|x)\mu^{\mathsf{X}}(\mathrm{d}x) \qquad \forall A \in \mathcal{X}, B \in \mathcal{F} \tag{143}$$

$$\iff \tag{144}$$

$$\mu = \tag{145}$$



and

$$\mu^{|\mathsf{X}}(\mathsf{X}^{-1}(A)|x) = \delta_x(A) \tag{146}$$

Lemma 6.8. *Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, if there is a regular proper conditional probability $\mu^{|\mathsf{X}} : X \to\!\!\!\!\bullet\ \Omega$ then there is a valid conditional distribution $\mu^{\mathsf{Y}|\mathsf{X}}$.*

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

Defining $\mathsf{O} := \mathrm{id}_\Omega$ (the identity function $\Omega \to \Omega$), $\mu^{\mathsf{I}\mathsf{X}}$ is a version of $\mu^{\mathsf{O}|\mathsf{X}}$. Note also that $\mathsf{Y} = \mathsf{Y} \circ \mathsf{O}$. Thus by Lemma 6.1, $\mathbb{K}$ is a version of $\mu^{\mathsf{Y}|\mathsf{X}}$.

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

$$\mathbb{K}(\mathsf{Y} \bowtie A|x) = \delta_x(\mathsf{Y}^{-1}(A)) \tag{147}$$
$$= 0 \tag{148}$$

Thus $\mathbb{K}$ is valid. □

Theorem 6.9 (Validity)**.** *Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathbb{J} \in \Delta(X)$ with Ω and X standard measurable, there exists some $\mu \in \Delta(\Omega)$ such that $\mu^{\mathsf{X}} = \mathbb{J}$ if and only if $\mathbb{J}$ is a valid distribution.*

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

$$\mu^{\mathsf{X}}(x) = \mu(\mathsf{X} \bowtie x) \tag{149}$$
$$= \sum_{x' \in X} \mathbb{J}(x')\mathbb{K}(\mathsf{X} \bowtie x|x') \tag{150}$$
$$= 0 \tag{151}$$
$$\neq \mathbb{J}(x) \tag{152}$$

□

Lemma 6.10 (Semidirect product defines an intersection of probability sets)**.** *Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to (X, \mathcal{X})$, $\mathsf{Y} : \Omega \to (Y, \mathcal{Y})$, $\mathsf{Z} : \Omega \to (Z, \mathcal{Z})$ all standard measurable and maximal probability sets $\mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}[M]}$ and $\mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{Y}\mathsf{X}[M]}$ then defining*

$$\mathbb{R}_{\{\}}^{\mathsf{Y}\mathsf{Z}|\mathsf{X}} := \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{Y}\mathsf{X}} \tag{153}$$

we have

$$\mathbb{R}_{\{\}}^{\mathsf{Y}\mathsf{Z}|\mathsf{X}[M]} = \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}[M]} \cap \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{Y}\mathsf{X}[M]} \tag{154}$$

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

$$\mathbb{R}_a^{\mathsf{X}\mathsf{Y}\mathsf{Z}} = \mathbb{R}_a^{\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{Y}\mathsf{X}} \tag{155}$$
$$\implies \mathbb{R}_a^{\mathsf{X}\mathsf{Y}} = \mathbb{R}_a^{\mathsf{X}} \odot \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \tag{156}$$
$$\wedge \mathbb{R}_a^{\mathsf{X}\mathsf{Y}\mathsf{Z}} = \mathbb{R}_a^{\mathsf{X}\mathsf{Y}} \odot \mathbb{Q}_{\{\}}^{\mathsf{Z}|\mathsf{Y}\mathsf{X}} \tag{157}$$

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

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

Lemma 6.11 (Equivalence of validity definitions)**.** *Given $\mathsf{X} : \Omega \to X$, with Ω and X standard measurable, a probability measure $\mathbb{P}^{\mathsf{X}} \in \Delta(X)$ is valid if and only if the conditional $\mathbb{P}^{\mathsf{X}|*} := * \mapsto \mathbb{P}^{\mathsf{X}}$ is valid.*

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

$$\forall A \in \mathcal{X} : \mathsf{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathsf{X}|*}(A|*) = 0 \tag{158}$$

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

$$\forall A \in \mathcal{X} : \mathsf{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathsf{X}}(A) = 0 \tag{159}$$

□

Lemma 6.12 (Semidirect product of valid candidate conditionals is valid)**.** *Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ (all spaces standard measurable) and any valid candidate conditional $\mathbb{P}^{\mathsf{Y|X}}$ and $\mathbb{Q}^{\mathsf{Z|YX}}$, $\mathbb{P}^{\mathsf{Y|X}} \odot \mathbb{Q}^{\mathsf{Z|YX}}$ is also a valid candidate conditional.*

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

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

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

$$\begin{aligned} \mathbb{R}^{\mathsf{YZ|X}}(B \times C|x) &= \int_B \mathbb{Q}^{\mathsf{Z|YX}}(C|y,x)\mathbb{P}^{\mathsf{Y|X}}(dy|x) && (160)\\ &\leq \mathbb{P}^{\mathsf{Y|X}}(B|x) && (161)\\ &= 0 && (162) \end{aligned}$$

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

$$\begin{aligned} \mathbb{R}^{\mathsf{YZ|X}}(B \times C|x) &= \int_B \mathbb{Q}^{\mathsf{Z|YX}}(C|y,x)\mathbb{P}^{\mathsf{Y|X}}(dy|x) && (163)\\ &= 0 && (164) \end{aligned}$$

□

Corollary 6.13 (Valid conditionals are validly extendable to valid distributions)**.** *Given Ω, $\mathsf{U} : \Omega \to U$, $\mathsf{W} : \Omega \to W$ and a valid conditional $\mathbb{T}^{\mathsf{W}|\mathsf{U}}$, then for any valid conditional $\mathbb{V}^{\mathsf{U}}$, $\mathbb{V}^{\mathsf{U}} \odot \mathbb{T}^{\mathsf{W}|\mathsf{U}}$ is a valid probability.*

Proof. Applying Lemma 6.12 choosing $\mathsf{X} = *$, $\mathsf{Y} = \mathsf{U}$, $\mathsf{Z} = \mathsf{W}$ and $\mathbb{P}^{\mathsf{Y}|\mathsf{X}} = \mathbb{V}^{\mathsf{U}|*}$ and $\mathbb{Q}^{\mathsf{Z}|\mathsf{YX}} = \mathbb{T}^{\mathsf{W}|\mathsf{U}*}$ we have $\mathbb{R}^{WU|*} := \mathbb{V}^{\mathsf{U}|*} \odot \mathbb{T}^{\mathsf{W}|\mathsf{U}*}$ is a valid conditional probability. Then $\mathbb{R}^{\mathsf{WU}} \cong \mathbb{R}^{\mathsf{WU}|*}$ is valid by Theorem 6.11. □

Theorem 6.14 (Validity of conditional probabilities)**.** *Suppose we have Ω, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, with Ω, X, Y discrete. A conditional $\mathbb{T}^{\mathsf{Y}|\mathsf{X}}$ is valid if and only if for all valid candidate distributions $\mathbb{V}^{\mathsf{X}}$, $\mathbb{V}^{\mathsf{X}} \odot \mathbb{T}^{\mathsf{Y}|\mathsf{X}}$ is also a valid candidate distribution.*

Proof. If: this follows directly from Corollary 6.13.

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

$$(\mathbb{V}^{\mathsf{X}} \odot \mathbb{T}^{\mathsf{Y}|\mathsf{X}})(x, y) = \mathbb{T}^{\mathsf{Y}|\mathsf{X}}(y|x)\mathbb{V}^{\mathsf{X}}(x) \tag{165}$$

$$= \mathbb{T}^{\mathsf{Y}|\mathsf{X}}(y|x) \tag{166}$$

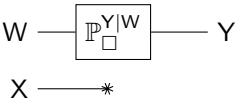
$$> 0 \tag{167}$$

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

6.3 Conditional independence

Theorem 6.15. *Given standard measurable Ω, a probability model $\mathbb{P}$ and variables $\mathsf{W} : \Omega \to W$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}} \mathsf{X}|\mathsf{W}$ if and only if there exists some version of $\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}$ and $\mathbb{P}^{\mathsf{Y}|\mathsf{W}}$ such that*

$$\mathbb{P}^{\mathsf{Y}|\mathsf{WX}} = \tag{168}$$



$$\iff \mathbb{P}^{\mathsf{Y}|\mathsf{WX}}(y|w, x) = \mathbb{P}^{\mathsf{Y}|\mathsf{W}}(y|w) \tag{169}$$

Proof. See Cho and Jacobs (2019). □

6.4 Extended conditional independence

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

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

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

Definition 6.16 (Extended conditional independence)**.** adf

Theorem 6.17. *Let* $\mathsf{A}^* = \mathsf{A} \circ \mathsf{V}$, $\mathsf{B}^* = \mathsf{B} \circ \mathsf{V}$, $\mathsf{C}^* = \mathsf{C} \circ \mathsf{V}$ *(*$(\mathsf{A}, \mathsf{B}, \mathsf{C})$ *are* $\mathcal{V}$*-measurable) and* $\mathsf{D}^* = \mathsf{D} \circ \mathsf{W}$,$\mathsf{E}^* = \mathsf{E} \circ \mathsf{W}$ *where* W *is discrete and* $\mathsf{W} = (\mathsf{D}^*, \mathsf{E}^*)$*. In addition, let* $\mathbb{P}_{\alpha}^{\mathsf{W}}$ *be some probability distribution on* W *such that* $w \in \mathsf{W}(\Omega) \implies \mathbb{P}_{\alpha}^{\mathsf{W}}(w) > 0$*. Then, denoting extended conditional independence with* $\perp\!\!\!\perp_{\mathbb{P},ext}$ *and* $\mathbb{P}_{\alpha}^{\mathsf{VW}} := \mathbb{P}_{\alpha}^{\mathsf{W}} \odot \mathbb{P}^{\mathsf{V}|\mathsf{W}}$

$$\mathsf{A} \perp\!\!\!\perp_{\mathbb{P},ext} (\mathsf{B}, \mathsf{D})|(\mathsf{C}, \mathsf{E}) \iff \mathsf{A}^* \perp\!\!\!\perp_{\mathbb{P}_\alpha} (\mathsf{B}^*, \mathsf{D}^*)|(\mathsf{C}^*, \mathsf{E}^*) \tag{170}$$

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

Theorem 6.18. *Let* $\mathsf{A}^* = \mathsf{A} \circ \mathsf{V}$, $\mathsf{B}^* = \mathsf{B} \circ \mathsf{V}$, $\mathsf{C}^* = \mathsf{C} \circ \mathsf{V}$ *(*$(\mathsf{A}, \mathsf{B}, \mathsf{C})$ *are* $\mathcal{V}$*-measurable) and* $\mathsf{D}^* = \mathsf{D} \circ \mathsf{W}$,$\mathsf{E}^* = \mathsf{E} \circ \mathsf{W}$ *where* V, W *are discrete and* $\mathsf{W} = (\mathsf{D}^*, \mathsf{E}^*)$*. Then letting* $\mathbb{P}_{\alpha}^{\mathsf{VW}} := \mathbb{P}_{\alpha}^{\mathsf{W}} \odot \mathbb{P}^{\mathsf{V}|\mathsf{W}}$

$$\mathsf{A} \perp\!\!\!\perp^{1}_{\mathbb{P},ext} (\mathsf{B}, \mathsf{D})|(\mathsf{C}, \mathsf{E}) \iff \mathsf{A}^* \perp\!\!\!\perp_{\mathbb{P}} (\mathsf{B}^*, \mathsf{D}^*)|(\mathsf{C}^*, \mathsf{E}^*) \tag{171}$$

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

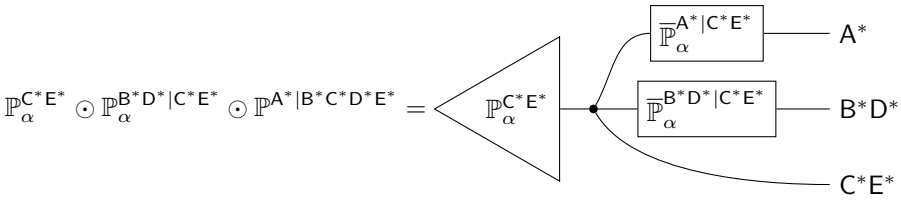
Only if:

For any β, $\mathbb{P}_{\beta}^{\mathsf{ABC}|\mathsf{DE}} = \mathbb{P}_{\beta}^{\mathsf{DE}} \odot \mathbb{P}^{\mathsf{ABC}|\mathsf{DE}}$. By Lemma 6.4, we have $\mathbb{P}^{\mathsf{A}|\mathsf{BCDE}}$ such that

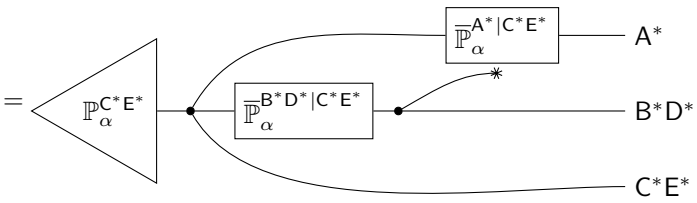
$$\begin{aligned}
\mathbb{P}_{\beta}^{\mathsf{A}^*\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} &= \mathbb{P}_{\beta}^{\mathsf{D}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{B}^*\mathsf{C}^*|\mathsf{D}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{A}^*|\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} && (172)\\
&= \mathbb{P}_{\beta}^{\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{A}^*|\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} && (173)\\
&= \mathbb{P}_{\beta}^{\mathsf{C}^*\mathsf{E}^*} \odot \mathbb{P}_{\beta}^{\mathsf{B}^*\mathsf{D}^*|\mathsf{C}^*\mathsf{E}^*} \odot \mathbb{P}^{\mathsf{A}^*|\mathsf{B}^*\mathsf{C}^*\mathsf{D}^*\mathsf{E}^*} && (174)
\end{aligned}$$

By Theorem 6.17, we have some α such that $\mathbb{P}_{\alpha}^{\mathsf{D}^*\mathsf{E}^*}$ is strictly positive on the range of $(\mathsf{D}^*, \mathsf{E}^*)$ and $\mathsf{A}^* \perp\!\!\!\perp_{\mathbb{P}_\alpha} (\mathsf{B}^*, \mathsf{D}^*)|(\mathsf{C}^*, \mathsf{E}^*)$.

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



(175)

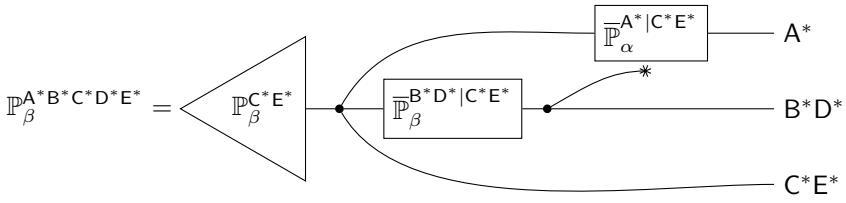


(176)

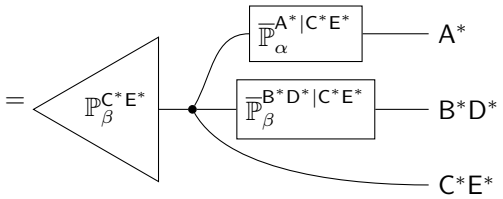
$$= \mathbb{P}_\alpha^{C^*E^*} \odot \mathbb{P}_\alpha^{B^*D^*|C^*E^*} \odot (\mathbb{P}_\alpha^{A^*|C^*E^*} \otimes \mathrm{erase}_{BD})$$

(177)

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



(178)



(179)

□

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

Definition 6.19 (Unresponsiveness)**.** Given discrete Ω, a probability gap model $\mathbb{P}_\square : A \to \Delta(\Omega)$, variables $\mathsf{W} : \Omega \to W$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, if there is some version of the conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}$ and $\mathbb{P}^{\mathsf{Y}|\mathsf{W}}_\square$ such that

$$\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}_\square = \quad \mathsf{W} \text{ — } \boxed{\mathbb{P}^{\mathsf{Y}|\mathsf{W}}_\square} \text{ — } \mathsf{Y} \qquad \mathsf{X} \text{ ——} * \tag{180}$$

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

Definition 6.20 (Domination)**.** Given a probability set $\mathbb{P}_{\{\}} \subset \Delta(\Omega)$, $\mathbb{P}_\alpha$ dominates $\mathbb{P}_{\{\}}$ if $\mathbb{P}_\beta(B) > 0 \implies \mathbb{P}_\alpha(B) > 0$ for all $\mathbb{P}_\beta in \mathbb{P}_{\{\}}$, $B \in \mathcal{F}$.

Theorem 6.21 (Conditional independence from kernel unresponsiveness)**.** *Given standard measurable Ω, variables $\mathsf{W} : \Omega \to W$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and a probability set $\mathbb{P}_{\{\}} : A \to \Delta(\Omega)$ with conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}_{\{\}}$ such that there is $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$ dominating $\mathbb{P}_{\{\}}$, $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{X}|\mathsf{W}$ if and only if there is a version of $\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}_{\{\}}$ unresponsive to W.*

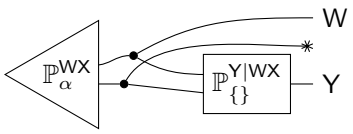
Proof. If: For every $\alpha \in A$ we can write

$$\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}_\alpha = \quad \mathsf{W} \text{ — } \boxed{\mathbb{P}^{\mathsf{Y}|\mathsf{W}}_\alpha} \text{ — } \mathsf{Y} \qquad \mathsf{X} \text{ ——} * \tag{181}$$

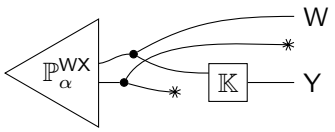
And so, by Theorem 6.15, $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{X}|\mathsf{W}$ for all $\alpha \in A$, and so $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{X}|\mathsf{W}$. Only if: For $\mathbb{P}_\alpha$ dominating $\mathbb{P}_{\{\}}$, by Theorem 6.15, there exists a version of $\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}_\alpha$ unresponsive to W. Because $\mathbb{P}_\alpha$ dominates $\mathbb{P}_{\{\}}$, $\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}_\alpha$ differs from $\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}_\beta$ on a set of measure 0 for any $\mathbb{P}_\beta \in \mathbb{P}_{\{\}}$, thus $\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}_\alpha$ is a version of $\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}_{\{\}}$ also. □

Corollary 6.22. *Given standard measurable Ω, variables $\mathsf{W} : \Omega \to W$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and a probability set $\mathbb{P}_{\{\}} : A \to \Delta(\Omega)$ with conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}_{\{\}}$, $\mathbb{P}^{\mathsf{Y}|\mathsf{W}}_{\{\}}$ exists if $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathsf{X}|\mathsf{W}$.*

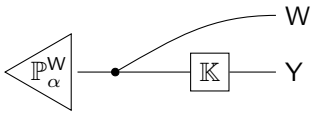
Proof. By Theorem 6.21, there is $\mathbb{K} : W \twoheadrightarrow Y$ such that for all α



$$\mathbb{P}_\alpha^{\mathsf{WY}} = \tag{182}$$



$$= \tag{183}$$



$$= \tag{184}$$

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.

Theorem 3.30. $[\forall x : (f(x) \implies g(x))] \implies [(\forall x : f(x)) \implies (\forall x : g(x))]$

Proof.

$$\forall x : f(x) \implies g(x) \quad \text{premise} \quad (185)$$

$$\forall x : f(x) \quad \text{premise} \quad (186)$$

$$f(a) \quad \text{UI on 186 sub } a/x \quad (187)$$

$$f(a) \implies g(a) \quad \text{UI on 185 sub } a/x \quad (188)$$

$$g(a) \quad \text{MP 187 and 188} \quad (189)$$

$$\forall x : g(x) \quad \text{UG on 189} \quad (190)$$

$$(\forall x : f(x)) \implies (\forall x : g(x)) \quad \text{CP } 186 - 190 \quad (191)$$

$$[\forall x : (f(x) \implies g(x))] \implies [(\forall x : f(x)) \implies (\forall x : g(x))] \quad \text{CP } 185\text{–}191 \quad (192)$$

Where UI: universal instantiation, UG: universal generalisation, MP: modus ponens and CP: conditoinal proof. With thanks to (https://math.stackexchange.com/users/252356/kylew) for the proof. □

6.5 Conclusion

Given a set of choices and the ability to compare the desirability of different outcomes, if we want to to compare the desirability of different choices then we need a function from choices to outcomes. If outcomes are to be represented probabilistically, we have proposed that we can represent the relevant kinds of functions using probability gap models, which are themselves defined using probability sets. Probability sets give us natural generalisations of well-established ideas of probabilistic variables, conditional probability and conditional independence, which we can make use of to reason about probabilistic models of choices and consequences.

Using this framework, we examine a particular question relevant to causal inference: when do "objective" collections of interventional distributions or distributions over potential outcomes exist? De Finetti previously addressed a similar question: when does an "objective" probability distribution describing a sequence of observations exist? He showed that under the assumption that the observations could be modeled exchangeably, an objective probability distribution appears as a parameter shared by a sequence of identically distributed observations, independent conditional on that parameter. We hypothesise that, generalising this argument to models with actions and responses, an "objective collection of interventional distributions" is a parameter shared by a conditionally independent and identical sequence of response conditionals.

Under this interpretation, we show that the existence of an "objective" response conditional is equivalent to the property of *causal contractibility* of a model of choices and outcomes. We discuss experiments where we thing causal contractibility might hold and experiments where we think it might not. The differences between the two can sometimes be subtle. This refines the idea put forward by Hernán (2016) that potential outcomes are well-defined when they are suitably precisely specified; in particular, we argue that the necessary kind of "precision" is that actions are deterministically specified when the decision maker's knowledge is consistent with a judgement of causal contractibility.

There are two challenges that arise when we try to apply this approach to typical causal inference problems. The first is that choice variables (that is, variables that represent a decision maker's choices) play a prominent role in our theory but in many common causal investigations they do not play such a role. Strictly speaking, conditional probability models may be appliccable to situations where no decision makers can be identified. However, they do seem to be a particularly natural fit for modelling the prospects a decision maker faces at the point of selecting a choice, and this interpretation played an important role in our investigation of the property of causal contractibility.

The second challenge, somewhat related to the first, is that we are often interested in causal investigations where the observed data are collected under somewhat different circumstances to the outcomes of actions. For example, observations might come from experiments conducted by another party with an action plan that is unknown to the decision maker.

A property of conditional probability models that may help bridge this gap is what we call *proxy control*. This is the condition where, given a sequence of experiments with choices D_i and outcomes Y_i causally contractibile with respect $(\mathsf{D}_i, \mathsf{Y}_i)$ pairs, if there exists some intermediate X_i such that $\mathsf{Y}_i \perp\!\!\!\perp \mathsf{D}_i | \mathsf{X}_i$ then causal contractibility also holds with respect to $(\mathsf{X}_i, \mathsf{Y}_i)$ pairs. This implies, for example, in a randomised experiemnt where the choices D_i are functions from a random source R_i to treatments X_i, we not only have response conditionals $\mathbb{P}_{\square}^{\mathsf{Y}_i|\mathsf{D}_i}$ that tell us how outcomes respond to treatment assignment functions, but also response conditionals $\mathbb{P}_{\square}^{\mathsf{Y}_i|\mathsf{X}_i}$ that tell us how outcomes respond to treatments.

The principle of proxy control is likely to be useful to analyse decision problems beyond idealised randomised experiments. For example, *causal inference by invariant prediction* (Peters et al., 2016) is a method of causal inference in which data is divided according to a number of different environments, characterised as "distributions observed under different interventions", and sets of variables that predict an outcome in the same manner in all environments are taken to be a sufficient set of causal ancestors fo the outcome. We speculate that, where causal inference by invariant prediction is possible, the situation can be modeled with a conditional probability model causally contractible with respect to (E, Y) where E is a variable representing the environment. Then, if we have $\mathsf{Y} \perp\!\!\!\perp \mathsf{E} | \mathsf{X}$, we also have causal contractibility with respect to (X, Y).

References

The Basic Symmetries. In Olav Kallenberg, editor, *Probabilistic Symmetries and Invariance Principles*, Probability and Its Applications, pages 24–68. Springer, New York, NY, 2005. ISBN 978-0-387-28861-1. doi: 10.1007/0-387-28861-9_2. URL https://doi.org/10.1007/0-387-28861-9_2.

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

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

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

George Boole. On the Theory of Probabilities. *Philosophical Transactions of the Royal Society of London*, 152:225–252, 1862. ISSN 0261-0523. URL https://www.jstor.org/stable/108830. Publisher: The Royal 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.

William Feller. *An Introduction to Probability Theory and its Applications, Volume 1*. J. Wiley & Sons: New York, 1968.

R.P. Feynman. *The Feynman lectures on physics*. Le cours de physique de Feynman. Intereditions, France, 1979. ISBN 978-2-7296-0030-3. INIS Reference Number: 13648743.

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

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

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

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

Miguel A. Hernán. Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*, 26(10):674–680, October 2016. ISSN 1047-2797. doi: 10.1016/j.annepidem.2016.08.016. URL http://www.sciencedirect.com/science/article/pii/S1047279716302800. Publisher: Elsevier.

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

Alan Hájek. What Conditional Probability Could Not Be. *Synthese*, 137(3): 273–323, December 2003. ISSN 1573-0964. doi: 10.1023/B:SYNT.0000004904.91112.16. URL https://doi.org/10.1023/B:SYNT.0000004904.91112.16.

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

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

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

Karl Menger. Random Variables from the Point of View of a General Theory of Variables. In Karl Menger, Bert Schweizer, Abe Sklar, Karl Sigmund, Peter Gruber, Edmund Hlawka, Ludwig Reich, and Leopold Schmetterer, editors, *Selecta Mathematica: Volume 2*, pages 367–381. Springer, Vienna, 2003. ISBN 978-3-7091-6045-9. doi: 10.1007/978-3-7091-6045-9_31. URL https://doi.org/10.1007/978-3-7091-6045-9_31.

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

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

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

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of*

the Royal Statistical Society: Series B (Statistical Methodology), 78(5):947–1012, 2016. ISSN 1467-9868. doi: 10.1111/rssb.12167. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12167.

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.

N. N. Vorobev. Consistent Families of Measures and Their Extensions. *Theory of Probability & Its Applications*, 7(2), 1962. doi: 10.1137/1107014. URL http://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=tvp&paperid=4710&option_lang=eng.

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

Appendix: