

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

David Johnston, Robert C. Williamson, Cheng Soon Ong

March 9, 2022

Contents

1	Introduction	2
1.1	Our approach	3
1.2	Contributions	5
2	Probability distributions, Markov kernels and string diagrams	6
2.0.1	Examples	8
2.0.2	Example: comb insertion	9
3	Variables in probabilistic models	10
3.1	Measurment procedures	11
3.2	Observable variables	13
3.3	Model variables	14
3.4	Variable sequences	14
3.5	Decision procedures	15
4	Decision problems	15
4.1	Other decision theoretic causal models	16
5	Probability sets	17
5.1	The roles of variables and probabilistic models	17
5.2	Standard probability theory	18
5.3	Not quite standard probability theory	19
5.4	Probability sets	20
5.5	Semidirect product and almost sure equality	21
5.6	Conditional independence	23
5.7	Uniform conditional independence	25
6	When do response conditionals exist?	28
6.1	Sequential decision models	28
6.2	Causal contractibility	30
6.3	Existence of response conditionals	34

6.4	Modelling different decision procedures	39
6.5	Decision procedures with response conditionals	41
6.6	Example: exchange commutativity in the context of treatment choices	42
6.7	Causal consequences of non-deterministic variables	45
6.8	Body mass index revisited	46
6.9	Inferred causal contractibility	47
7	Conclusion	48
7.1	Choices aren't always known	49
8	Appendix, needs to be organised	50
8.1	Existence of conditional probabilities	50
8.2	Validity	54
8.3	Conditional independence	57
8.4	Maximal probability sets and valid conditionals	59

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 and Taubman’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 applicable 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 unknown probability \mathbb{Q} , a “fixed but unknown causal effect $\mathbb{Q}(Y|do(X))$ ” 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 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. 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.

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.

Roadmap

2 Probability distributions, Markov kernels and string diagrams

This section needs updating to bring it into line with the current document

We make use of a string diagram notation for probabilistic reasoning. Graphical models are often employed in causal reasoning, and string diagrams are a particularly rigorous graphical notation for probabilistic models. It comes from the study of Markov categories. Markov categories are abstract categories that represent models of the flow of information. We can form Markov categories from collections of sets – for example, discrete sets or standard measurable sets – along with the Markov kernel product as the composition operation. Markov categories come equipped with a graphical language of *string diagrams*, and a coherence theorem which states that valid proofs using string diagrams correspond to valid theorems in *any* Markov category (Selinger, 2011). More comprehensive introductions to Markov categories can be found in Fritz (2020); Cho and Jacobs (2019). Thus, while we limit ourselves to discrete sets in this paper, any derivation that uses only string diagrams is more broadly applicable.

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

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

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

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

Definition 2.2 (Markov kernel (discrete sets)). A Markov kernel $\mathbb{K} : X \rightarrow Y$ is a matrix $(\mathbb{K}(y|x))_{x \in X, y \in Y} \in [0, 1]^{|X||Y|}$ such that $\sum_{y \in Y} \mathbb{K}(y|x) = 1$ for all $x \in X$. For $B \subset Y$ define $\mathbb{K}(B|x) = \sum_{y \in B} \mathbb{K}(y|x)$.

In the graphical language, Markov kernels are drawn as boxes with input and output wires, and probability measures (which are kernels with the domain $\{*\}$) are represented by triangles:

$$\mathbb{K} := \text{---} \boxed{\mathbb{K}} \text{---} \quad (1)$$

$$\mathbb{P} := \text{---} \triangleleft \boxed{\mathbb{P}} \text{---} \quad (2)$$

Two Markov kernels $\mathbb{L} : X \rightarrow Y$ and $\mathbb{M} : Y \rightarrow Z$ have a product $\mathbb{LM} : X \rightarrow Z$, given in the discrete case by the matrix product $\mathbb{LM}(z|x) = \sum_{y \in Y} \mathbb{M}(z|y)\mathbb{L}(y|x)$. Graphically, we represent products between compatible Markov kernels by joining wires together:

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

The Cartesian product $X \times Y := \{(x, y) | x \in X, y \in Y\}$. Given kernels $\mathbb{K} : W \rightarrow Y$ and $\mathbb{L} : X \rightarrow Z$, the tensor product $\mathbb{K} \otimes \mathbb{L} : W \times X \rightarrow Y \times Z$ given by $(\mathbb{K} \otimes \mathbb{L})(y, z | w, x) := K(y|w)L(z|x)$. The tensor product is graphically represented by drawing kernels in parallel:

$$\mathbb{K} \otimes \mathbb{L} := \begin{array}{c} W \text{---} \boxed{\mathbb{K}} \text{---} Y \\ X \text{---} \boxed{\mathbb{L}} \text{---} Z \end{array} \quad (4)$$

We read diagrams from left to right (this is somewhat different to Fritz (2020); Cho and Jacobs (2019); Fong (2013) but in line with Selinger (2011)), and any diagram describes a set of nested products and tensor products of Markov kernels. There are a collection of special Markov kernels for which we can replace the generic “box” of a Markov kernel with a diagrammatic elements that are visually suggestive of what these kernels accomplish.

The identity map $\text{id}_X : X \rightarrow X$ defined by $(\text{id}_X)(x'|x) = \llbracket x = x' \rrbracket$, where the Iverson bracket $\llbracket \cdot \rrbracket$ evaluates to 1 if \cdot is true and 0 otherwise, is a bare line:

$$\text{id}_X := X \text{---} X \quad (5)$$

We choose a particular 1-element set $\{*\}$ that acts as the identity in the sense that $\{*\} \times A \cong A \times \{*\} \cong A$ for any set A . The erase map $\text{del}_X : X \rightarrow \{*\}$ defined by $(\text{del}_X)(*|x) = 1$ is a Markov kernel that “discards the input”. It is drawn as a fuse:

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

The copy map $\text{copy}_X : X \rightarrow X \times X$ defined by $(\text{copy}_X)(x', x'' | x) = \llbracket x = x' \rrbracket \llbracket x = x'' \rrbracket$ is a Markov kernel that makes two identical copies of the input. It is drawn as a fork:

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

The swap map $\text{swap}_{X,Y} : X \times Y \rightarrow Y \times X$ defined by $(\text{swap}_{X,Y})(y', x' | x, y) = \llbracket x = x' \rrbracket \llbracket y = y' \rrbracket$ swaps two inputs, and is represented by crossing wires:

$$\text{swap}_X := \text{X} \quad (8)$$

Because we anticipate that the graphical notation will be unfamiliar, we will include some examples in the next section.

2.0.1 Examples

When translating string diagram notation to integral notation, a number of identities can speed up the process.

For arbitrary $\mathbb{K} : X \times Y \rightarrow Z$, $\mathbb{L} : W \rightarrow Y$

$$[(\text{id}_X \otimes \mathbb{L})\mathbb{K}](A|x, w) = \int_Y \int_X \mathbb{K}(z|x', y') \mathbb{L}(dy'|w) \text{id}_X(dx'|x) \quad (9)$$

$$= \int_Y \mathbb{K}(z|x, y') \mathbb{L}(dy'|w) \quad (10)$$

That is, an identity map passes its input to the next kernel in the product.

For arbitrary $\mathbb{K} : X \times Y \times Y \rightarrow Z$ (where we apply the above shorthand in the first line):

$$[(\text{id}_X \otimes \text{copy}_Y)\mathbb{K}](A|x, y) = \int_Y \int_Y \mathbb{K}(A|x, y', y'') \text{copy}_Y(dy' \times dy''|y) \quad (11)$$

$$= \mathbb{K}(A|x, y, y) \quad (12)$$

That is, the copy map passes along two copies of its input to the next kernel in the product.

For a collection of kernels $\mathbb{K}^n : Y^n \rightarrow Z$, $n \in [n]$, define $(y)^n = (y|i \in [n])$ and:

$$\text{copy}_Y^n := \begin{cases} \text{copy}_Y^{n-1}(\text{id}_{Y^{n-2}} \otimes \text{copy}_Y) & n > 2 \\ \text{copy}_Y & n = 2 \end{cases} \quad (13)$$

$$(\text{copy}_Y^2 \mathbb{K}^2)(z|y) = \mathbb{K}^2(z|y, y) \quad (14)$$

$$(15)$$

Suppose for induction

$$(\text{copy}_Y^{n-1} \mathbb{K}^{n-1})(z|y) = \mathbb{K}^{n-1}(z|(y)^{n-1}) \quad (16)$$

then

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

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

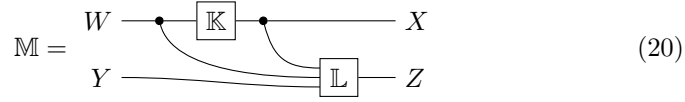
$$= \mathbb{K}^n(z|(y)^n) \quad (19)$$

That is, we can define the n -fold copy map that passes along n copies of its input to the next kernel in the product.

2.0.2 Example: comb insertion

The following examples illustrate 2-combs and the insertion operation, both of which we will define later. As an example in translating diagrams, we show how the diagrams for a 2-comb and 2-comb with an inserted Markov kernel can be translated to integral notation.

Consider the Markov kernels $\mathbb{K} : W \rightarrow X$, $\mathbb{L} : X \times W \times Y \rightarrow Z$ and the 2-comb $\mathbb{M} : W \times Y \rightarrow X \times Z$ defined as



Following the rules above, we can translate this to ordinary notation by first breaking it down into products and tensor products, and then evaluating these products

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

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

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

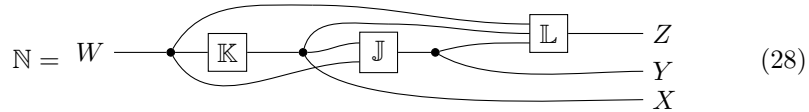
$$(\text{id}_X \otimes \mathbb{L})](A \times B|w, w, y) \quad (24)$$

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

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

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

If we are given additionally $\mathbb{J} : X \times W \rightarrow Y$, we can define a new Markov kernel $\mathbb{N} : W \rightarrow Z$ given by “inserting” \mathbb{J} into \mathbb{M} :



We can translate Equation 28 to

$$\mathbb{N}(A \times B \times C|w) = [\text{copy}_W(\mathbb{K}\text{copy}_Y^3 \otimes \text{id}_W)] \quad (29)$$

$$(\text{id}_Y \otimes \mathbb{J} \otimes \text{id}_Y)(\text{id}_Y \otimes \text{copy}_X \otimes \text{id}_Y) \quad (30)$$

$$(\mathbb{L} \otimes \text{id}_X \otimes \text{id}_Y)](A \times B \times C|w) \quad (31)$$

$$=[(\mathbb{K}\text{copy}_Y^3 \otimes \text{id}_W)(\text{id}_Y \otimes \mathbb{J} \otimes \text{id}_Y) \quad (32)$$

$$(\text{id}_Y \otimes \text{copy}_X \otimes \text{id}_Y) \quad (33)$$

$$(\mathbb{L} \otimes \text{id}_X \otimes \text{id}_Y)](A \times B \times C|w, w) \quad (34)$$

$$= \int_X \int_Y \mathbb{L}(C|x', w, y') \text{id}_X(A|x') \text{id}_Y(B|y') \mathbb{J}(dy'|x', w) \mathbb{K}(dx'|w) \quad (35)$$

$$= \int_A \int_B \mathbb{L}(C|x', w, y') \mathbb{J}(dy'|x', w) \mathbb{K}(dx'|w) \quad (36)$$

3 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 and 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 measurement 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 contains both observable 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 $F = MA$. 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 what 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 .

3.1 Measurement procedures

Definition 3.1 (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 3.2 (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 $\bigvee_{x \in A} \mathcal{B} \bowtie x$.

Definition 3.3 (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 measurement 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 3.4 (Composition of functions with procedures). Given a procedure \mathcal{B} that takes values in some set B , and a function $f : B \rightarrow C$, define the “composition” $f \circ \mathcal{B}$ to be any procedure \mathcal{C} that yields $f(x)$ whenever \mathcal{B} yields x . We can construct such a procedure by describing the steps: first, do \mathcal{B} and secondly, apply f to the value yielded by \mathcal{B} .

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

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

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

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

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

One might wonder whether there is also some kind of “append” operation that takes a standalone \mathcal{M} and a standalone \mathcal{A} and returns a procedure $(\mathcal{M}, \mathcal{A})$. Unlike function composition, this would be an operation that acts on two procedures rather than a procedure and a function. 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 (Ψ, \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 (Ψ, \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.

3.2 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 3.5 (Observable variable). Given a measurement procedure \mathcal{S} taking values in (Ψ, \mathcal{E}) , an observable variable is a pair $(X \circ \mathcal{S}, X)$ where $X : (\Psi, \mathcal{E}) \rightarrow (X, \mathcal{X})$ is a measurable function and $\mathcal{X} := X \circ \mathcal{S}$ is the measurement procedure induced by X and \mathcal{S} .

For the model $F = MA$, for example, suppose we have a complete 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}, F) where $\mathcal{F} := F \circ \mathcal{S}$ and $F : X \times Y \times Z \rightarrow X$ 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 complete procedure here. Given the complete procedure \mathcal{S} , a variable $X : \Psi \rightarrow X$ and the corresponding procedure $\mathcal{X} = X \circ \mathcal{S}$, the proposition “ \mathcal{X} yields x ” is equivalent to the proposition “ \mathcal{S} yields a value in $X^{-1}(x)$ ”. Because of this, we define the *event* $X \bowtie x$ to be the set $X^{-1}(x)$.

Definition 3.6 (Event). Given the complete procedure \mathcal{S} taking values in Ψ and an observable variable $(X \circ \mathcal{S}, X)$ for $X : \Psi \rightarrow X$, the *event* $X \bowtie x$ is the set $X^{-1}(x)$ for any $x \in X$.

If we are given an observation “ \mathcal{X} yields x ”, then the corresponding event $X \bowtie x$ 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 $Y = y$ already has a meaning, namely that Y is a constant function everywhere equal to y .

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

3.3 Model variables

Observable variables are special in the sense that they are tied to a particular measurement procedure \mathcal{S} . However, the measurement 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 complete measurement procedure \mathcal{S} , retaining only its set of possible values (Ψ, \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 (Ω, \mathcal{F}) that does not correspond to the possible values that \mathcal{S} might yield. In that case, we require a surjective model variable $S : \Omega \rightarrow \Psi$ called the complete observable variable, and every observable variable $(X' \circ \mathcal{S}, X')$ is associated with the model variable $X := X' \circ S$.

An *unobserved variable* is a variable whose set of possible values is not constrained by the results of the measurement procedure.

Definition 3.7 (Unobserved variable). Given a sample space (Ω, \mathcal{F}) and a complete observable variable $S : \Omega \rightarrow \Psi$, a model variable $Y : \Omega \rightarrow Y$ is *unobserved* if $Y(S \bowtie s) = Y$ for all $s \in \Psi$.

3.4 Variable sequences

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

3.5 Decision procedures

Our central problems are those in which we aim to decide on one choice from a set of possible choices, and this involves comparing the consequences that we expect to arise from each choice. A basic principle we adopt is that models are informed by the measurement procedure – the question of whether or not a mathematical model is appropriate depends on the measurement procedure it is modelling. We do not prescribe how this dependency plays out, but we do hold that one cannot decide a mathematical model to be appropriate in the absence of a description of the measurement procedure.

Putting both of these together, this means that in order to find an appropriate model of a decision problem we need a description of a measurement procedure for each possible choice. We could in principle describe a single measurement procedure that first determines the outcome of the decision, and then for each potential choice specifies how to conduct the rest of the procedure. However, we can often make decisions without including the decision making process in the model, so avoiding this extra information saves a substantial amount of complication.

We avoid this problem by specifying less complete measurement procedures. A *decision procedure* is a collection of measurement procedures, one for each element of a set of potential choices A . We have a background understanding – and maybe even a precise algorithm – for deciding on an element of A , but we leave this out of our model of consequences.

4 Decision problems

We want to construct models to help make decisions. For our purposes, “making a decision” means choosing some element of a mathematically well-defined set $\alpha \in C$, and following a measurement procedure S_α associated with the choice $\alpha \in C$ (see Section 3.5). We suppose that each S_α is modeled by a probability model \mathbb{P}_α on a shared sample space (Ω, \mathcal{F}) . Decision making also involves comparing the outcomes of different choices (that is, comparing the probability models \mathbb{P}_α associated with each choice) and selecting one of the “best” decisions, but we leave questions of comparison in the background.

The way we treat consequences of decisions is, in a sense, the opposite of the way we treat conducting measurements. A measurement involves some unclear measurement procedure that interacts with the world and leaves us with a collection of well-defined mathematical objects. Our view of the consequences of making a decision, in contrast, is that we assume that we start with some element of a well-defined set C which is then mapped to some unclear measurement procedure. If a measurement is a “function” whose domain is actions in the world, the consequences of a decision is a “function” whose codomain is actions in the world.

We make the assumption that each choice is associated with a measurement procedure S_α modeled by probability distribution \mathbb{P}_α . This is a Bayesian

approach – uncertainty over the outcomes of a measurement procedure is represented with a single probability measure. It is not our intention to suggest that this is the only way of representing uncertain knowledge, and it may be interesting to extend our theory to other methods for representing uncertain outcomes of a measurement procedure. A particularly simple extension would be to model each \mathcal{S}_α with a probability set rather than a single probability distribution.

The model of a decision problem of this form is naturally captured by a probability set $\mathbb{P}_C := \{\mathbb{P}_\alpha | \alpha \in C\}$.

4.1 Other decision theoretic causal models

There have been a number of formalisations of decision theoretic foundations of causal inference. All share the feature that there is a basic set of choices/interventions/regimes that may be chosen from, and a probability distribution is associated with each element of this set, so they all induce probability sets.

Lattimore and Rohde (2019a) and Lattimore and Rohde (2019b) describe a method for reformulating causal Bayesian networks as a set of probability distributions indexed by an intervention set T . Their algorithm *CausalBayesConstruct* is a method for translating directly from causal Bayesian networks with a specification of interventions to probability sets.

A key feature of the *CausalBayesConstruct* algorithm is that every probability distribution in the set can be represented as a product of the same set of conditional probabilities - clearly, these must be uniform conditional probabilities. We posit, therefore, that d-separation in probabilistic graphical models corresponds to *uniform conditional independence* given in Definition 5.29, on the basis of Theorem 5.31.

An alternative decision theoretic foundation has been developed in Dawid (2021, 2010, 2000). A key contribution of this literature is the notion of extended conditional independence (formally described in Constantinou and Dawid (2017), see Section 5.7), and its application to probability sets. Many common causal models have been described as probability sets in which certain extended conditional independence statements hold. A second contribution of Dawid (2021) is to develop a lower level justification for the use of probability sets modelling “generic variables” that appears in earlier work. Our work is an extension of this latter investigation.

Heckerman and Shachter (1995) also explore a decision theoretic approach to causal inference. Their approach differs from the previous two in two ways: first, they posit a set of choices and a set of unobserved states, and consider models that map $\text{States} \times \text{Choices} \rightarrow \text{Outcomes}$, instead of mapping choices only to outcomes. Secondly, they consider only deterministic maps rather than general probability distribution valued maps. This approach is based on the decision theory of Savage (1954). They consider an alternative “conditional independence-like” property of these models that they call *limited unresponsiveness*.

5 Probability sets

5.1 The roles of variables and probabilistic models

The sample space (Ω, \mathcal{F}) along with the measurement procedure(s) \mathcal{S} and the associated model variable \mathbf{S} is a “model skeleton”. The criterion of *compatibility with observation* establishes a relation between the results of measurements 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 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 (Ω, \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 applicable to reasoning 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 independence, 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.

5.2 Standard probability theory

Definition 5.1 (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 5.2 (Probability measure). Given a measurable space (X, \mathcal{X}) , a probability measure is a σ -additive function $\mu : \mathcal{X} \rightarrow [0, 1]$ 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 5.3 (Markov kernel). Given measurable spaces (X, \mathcal{X}) and (Y, \mathcal{Y}) , a Markov kernel $\mathbb{Q} : X \rightarrow Y$ is a map $Y \times \mathcal{X} \rightarrow [0, 1]$ 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 5.4 (Delta measure). Given a measurable space (X, \mathcal{X}) and $x \in X$, $\delta_x \in \Delta(X)$ is the measure defined by $\delta_x(A) := \mathbb{I}[x \in A]$ for all $A \in \mathcal{X}$

Definition 5.5 (Probability space). A probability space is a triple $(\mu, \Omega, \mathcal{F})$, where μ is a base measure on \mathcal{F} and (Ω, \mathcal{F}) is a measurable space.

Definition 5.6 (Variable). Given a measurable space (Ω, \mathcal{F}) and a measurable space of values (X, \mathcal{X}) , an X -valued variable is a measurable function $X : (\Omega, \mathcal{F}) \rightarrow (X, \mathcal{X})$.

Definition 5.7 (Sequence of variables). Given a measurable space (Ω, \mathcal{F}) and two variables $X : (\Omega, \mathcal{F}) \rightarrow (X, \mathcal{X})$, $Y : (\Omega, \mathcal{F}) \rightarrow (Y, \mathcal{Y})$, $(X, Y) : \Omega \rightarrow X \times Y$ is the variable $\omega \mapsto (X(\omega), Y(\omega))$.

Definition 5.8 (Marginal distribution with respect to a probability space). Given a probability space $(\mu, \Omega, \mathcal{F})$ and a variable $X : \Omega \rightarrow (X, \mathcal{X})$, we can define the *marginal distribution* of X with respect to μ , $\mu^X : \mathcal{X} \rightarrow [0, 1]$ by $\mu^X(A) := \mu(X^{-1}(A))$ for any $A \in \mathcal{X}$.

Definition 5.9 (Distribution-kernel products). Given (X, \mathcal{X}) , (Y, \mathcal{Y}) a probability distribution $\mu \in \Delta(X)$ and a Markov kernel $\mathbb{K} : X \rightarrow Y$, $\mu\mathbb{K}$ is a probability distribution on (Y, \mathcal{Y}) defined by

$$\mu\mathbb{K}(A) := \int_X \mathbb{K}(A|x)\mu(dx) \quad (41)$$

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

Definition 5.10 (Kernel-kernel products). Given (X, \mathcal{X}) , (Y, \mathcal{Y}) , (Z, \mathcal{Z}) and Markov kernels $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : Y \rightarrow Z$, \mathbb{KL} is a Markov kernel $X \rightarrow Z$ defined by

$$\mathbb{KL}(A|x) := \int_Y \mathbb{L}(A|y) \mathbb{K}(dy|x) \quad (42)$$

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

Lemma 5.11 (Marginal distribution as a kernel product). *Given a probability space $(\mu, \Omega, \mathcal{F})$ and a variable $\mathbf{X} : \Omega \rightarrow (X, \mathcal{X})$, define $\mathbb{F}_{\mathbf{X}} : \Omega \rightarrow X$ by $\mathbb{F}_{\mathbf{X}}(A|\omega) = \delta_{\mathbf{X}(\omega)}(A)$, then*

$$\mu^{\mathbf{X}} = \mu \mathbb{F}_{\mathbf{X}} \quad (43)$$

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

$$\mu \mathbb{F}_{\mathbf{X}}(A) = \int_{\Omega} \delta_{\mathbf{X}(\omega)}(A) d\mu(\omega) \quad (44)$$

$$= \int_{\mathbf{X}^{-1}(\omega)} d\mu(\omega) \quad (45)$$

$$= \mu^{\mathbf{X}}(A) \quad (46)$$

□

5.3 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 5.12 (Probability measures as Markov kernels). Given a measurable space (X, \mathcal{X}) and $\mu \in \Delta(X)$, the Markov kernel $\mathbb{K} : \{*\} \rightarrow X$ 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 5.13 (Conditional distribution). Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $\mathbf{X} : \Omega \rightarrow X$, $\mathbf{Y} : \Omega \rightarrow Y$, the probability of \mathbf{Y} given \mathbf{X} is any Markov

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

$$\mu^{XY}(A \times B) = \int_A \mu^{Y|X}(B|x) d\mu^X(x) \quad \forall A \in \mathcal{X}, B \in \mathcal{Y} \quad (47)$$

$$\iff \quad (48)$$

$$\mu^{XY} = \begin{array}{c} \text{X} \\ \nearrow \\ \triangleleft \mu^X \\ \bullet \\ \boxed{\mu^{Y|X}} \longrightarrow \text{Y} \end{array} \quad (49)$$

We define higher order conditionals as “conditionals of conditionals”.

Definition 5.14 (Higher order conditionals). Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ and $Z : \Omega \rightarrow Z$, a higher order conditional $\mu^{Z|(Y|X)} : X \times Y \rightarrow Z$ is any Markov kernel such that, for some $\mu^{Y|X}$,

$$\mu^{ZY|X}(B \times C|x) = \int_B \mu^{Z|(Y|X)}(C|x, y) \mu^{Y|X}(dy|x) \quad (50)$$

$$\iff \quad (51)$$

$$\mu^{ZY|X} = \begin{array}{c} \text{Y} \\ \nearrow \\ \boxed{\mu^{Y|X}} \longrightarrow \bullet \\ \nearrow \quad \searrow \\ \boxed{\mu^{Z|(Y|X)}} \longrightarrow \text{Z} \\ \nwarrow \\ \text{X} \end{array} \quad (52)$$

Higher order conditionals are useful because $\mu^{Z|(Y|X)}$ is a version of $\mu^{Z|YX}$, so if we're given $\mu^{ZY|X}$ but not μ itself, we use the higher order conditional $\mu^{Z|(Y|X)}$ as a version of $\mu^{X|YX}$. This also hold for conditional with respect to probability sets, which we will introduce later (Theorem 8.4).

Furthermore, given $\mu^{XY|Z}$ and X, Y standard measurable, it has recently been proven that a higher order conditional $\mu^{Z|(Y|X)}$ exists Bogachev and Malofeev (2020), Theorem 3.5. See also Theorem 8.3 for the extension of this theorem to probability sets.

5.4 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 respect 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 5.15 (Probability set). A probability set $\mathbb{P}_{\{\}}$ on (Ω, \mathcal{F}) is a collection of probability measures on (Ω, \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 5.8 and 5.13 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 5.21) with respect to probability sets.

Definition 5.16 (Marginal probability with respect to a probability set). Given a sample space (Ω, \mathcal{F}) , a variable $X : \Omega \rightarrow X$ and a probability set $\mathbb{P}_{\{\}}$, the marginal distribution $\mathbb{P}_{\{\}}^X = \mathbb{P}_{\alpha}^X$ for any $\mathbb{P}_{\alpha} \in \mathbb{P}_{\{\}}$ if a distribution satisfying this condition exists. Otherwise, it is undefined.

Definition 5.17 (Uniform conditional distribution with respect to a probability set). Given a sample space (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$ and a probability set $\mathbb{P}_{\{\}}$, a uniform conditional distribution $\mathbb{P}_{\{\}}^{Y|X}$ is any Markov kernel $X \rightarrow Y$ such that $\mathbb{P}_{\{\}}^{Y|X}$ is an $Y|X$ conditional probability of \mathbb{P}_{α} for all $\mathbb{P}_{\alpha} \in \mathbb{P}_{\{\}}$. If no such Markov kernel exists, $\mathbb{P}_{\{\}}^{Y|X}$ is undefined.

Definition 5.18 (Uniform higher order conditional distribution with respect to a probability set). Given a sample space (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ and $Z : \Omega \rightarrow Z$ and a probability set $\mathbb{P}_{\{\}}$, if $\mathbb{P}_{\{\}}^{ZY|X}$ exists then a uniform higher order conditional $\mathbb{P}_{\{\}}^{Z|(Y|X)}$ is any Markov kernel $X \times Y \rightarrow Z$ that is a higher order conditional of some version of $\mathbb{P}_{\{\}}^{ZY|X}$. If no $\mathbb{P}_{\{\}}^{ZY|X}$ exists, $\mathbb{P}_{\{\}}^{Z|(Y|X)}$ is undefined.

Under the assumption of standard measurable spaces, the existence of a uniform conditional distribution $\mathbb{P}_{\{\}}^{ZY|X}$ implies the existence of a higher order conditional $\mathbb{P}_{\{\}}^{Z|(Y|X)}$ with respect to the same probability set (Theorem 8.3). $\mathbb{P}_{\{\}}^{Z|(Y|X)}$ is in turn a version of the uniform conditional distribution $\mathbb{P}_{\{\}}^{Z|YX}$ (Theorem 8.4). Thus, from the existence of $\mathbb{P}_{\{\}}^{ZY|X}$ we can derive the existence of $\mathbb{P}_{\{\}}^{Z|YX}$.

5.5 Semidirect product and almost sure equality

The operation used in Equation 49 that combines μ^X and $\mu^{Y|X}$ 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 : $\mathbb{K} \stackrel{\mu^X}{\cong} \mathbb{L}$ if $\mu^X \odot \mathbb{K} = \mu^X \odot \mathbb{L}$ (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 5.19 (Semidirect product). Given $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : Y \times X \rightarrow Z$, define the copy-product $\mathbb{K} \odot \mathbb{L} : X \rightarrow Y \times Z$ as

$$\mathbb{K} \odot \mathbb{L} := \text{copy}_X(\mathbb{K} \otimes \text{id}_X)(\text{copy}_Y \otimes \text{id}_X)(\text{id}_Y \otimes \mathbb{L}) \quad (53)$$

$$= \begin{array}{c} \text{Diagram: } X \text{ splits into two paths. The top path goes through box } \mathbb{K} \text{ to } Y. \text{ The bottom path goes through box } \mathbb{L} \text{ to } Z. \end{array} \quad (54)$$

$$\iff \quad (55)$$

$$(\mathbb{K} \odot \mathbb{L})(A \times B|x) = \int_A \mathbb{L}(B|y, x) \mathbb{K}(dy|x) \quad A \in \mathcal{Y}, B \in \mathcal{Z} \quad (56)$$

Lemma 5.20 (Semidirect product is associative). Given $\mathbb{K} : X \rightarrow Y$, $\mathbb{L} : Y \times X \rightarrow Z$ and $\mathbb{M} : Z \times Y \times X \rightarrow W$

$$(\mathbb{K} \odot \mathbb{L}) \odot \mathbb{M} = \mathbb{K} \odot (\mathbb{L} \odot \mathbb{M}) \quad (57)$$

$$(58)$$

Proof.

$$(\mathbb{K} \odot \mathbb{L}) \odot \mathbb{M} = \begin{array}{c} \text{Diagram: } X \text{ splits into two paths. The top path goes through box } \mathbb{K} \text{ to } Y. \text{ The bottom path goes through box } \mathbb{L} \text{ to } Z. \text{ Then } Y \text{ and } Z \text{ split into two paths. The top path goes through box } \mathbb{M} \text{ to } W. \end{array} \quad (59)$$

$$= \begin{array}{c} \text{Diagram: } X \text{ splits into two paths. The top path goes through box } \mathbb{K} \text{ to } Y. \text{ The bottom path goes through box } \mathbb{L} \text{ to } Z. \text{ Then } Y \text{ and } Z \text{ split into two paths. The top path goes through box } \mathbb{M} \text{ to } W. \end{array} \quad (60)$$

$$= \mathbb{K} \odot (\mathbb{L} \odot \mathbb{M}) \quad (61)$$

□

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

Definition 5.21 (Almost sure equality). Two Markov kernels $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \rightarrow Y$ are almost surely equal $\stackrel{\mathbb{P}_{\{\}}}{\cong}$ with respect to a probability set $\mathbb{P}_{\{\}}$ and variable $X : \Omega \rightarrow X$ if for all $\mathbb{P}_{\alpha} \in \mathbb{P}_{\{\}}$,

$$\mathbb{P}_{\alpha}^X \odot \mathbb{K} = \mathbb{P}_{\alpha}^X \odot \mathbb{L} \quad (62)$$

Lemma 5.22 (Uniform conditional distributions are almost surely equal). If $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \rightarrow Y$ are both versions of $\mathbb{P}_{\{\}}^{Y|X}$ then $\mathbb{K} \stackrel{\mathbb{P}_{\{\}}}{\cong} \mathbb{L}$

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

$$\mathbb{P}_\alpha^X \odot \mathbb{K} = \mathbb{P}_\alpha^{XY} \quad (63)$$

$$= \mathbb{P}_\alpha^X \odot \mathbb{L} \quad (64)$$

□

Lemma 5.23 (Substitution of almost surely equal Markov kernels). *Given \mathbb{P}_Ω , if $\mathbb{K} : X \times Y \rightarrow Z$ and $\mathbb{L} : X \times Y \rightarrow Z$ are almost surely equal $\mathbb{K} \stackrel{\mathbb{P}_\Omega}{\cong} \mathbb{L}$, then for any $\mathbb{P}_\alpha \in \mathbb{P}_\Omega$*

$$\mathbb{P}_\alpha^{Y|X} \odot \mathbb{K} \stackrel{\mathbb{P}_\Omega}{\cong} \mathbb{P}_\alpha^{Y|X} \odot \mathbb{L} \quad (65)$$

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

$$\mathbb{P}_\alpha^{XY} \odot \mathbb{K} = (\mathbb{P}_\alpha^X \odot \mathbb{P}_\Omega^{Y|X}) \odot \mathbb{K} \quad (66)$$

$$= \mathbb{P}_\alpha^X \odot (\mathbb{P}_\Omega^{Y|X} \odot \mathbb{K}) \quad (67)$$

$$= \mathbb{P}_\alpha^X \odot (\mathbb{P}_\Omega^{Y|X} \odot \mathbb{L}) \quad (68)$$

□

Theorem 5.24 (Semidirect product of uniform conditional distributions is a joint uniform conditional distribution). *Given a probability set \mathbb{P}_Ω on (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ and uniform conditional distributions $\mathbb{P}_\Omega^{Y|X}$ and $\mathbb{P}_\Omega^{Z|XY}$, then $\mathbb{P}_\Omega^{YZ|X}$ exists and is equal to*

$$\mathbb{P}_\Omega^{YZ|X} = \mathbb{P}_\Omega^{Y|X} \odot \mathbb{P}_\Omega^{Z|XY} \quad (69)$$

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

$$\mathbb{P}_\alpha^{XYZ} = \mathbb{P}_\alpha^X \odot \mathbb{P}_\alpha^{YZ|X} \quad (70)$$

$$= \mathbb{P}_\alpha^X \odot (\mathbb{P}_\alpha^{Y|X} \odot \mathbb{P}_\alpha^{Z|YX}) \quad (71)$$

$$= \mathbb{P}_\alpha^X \odot (\mathbb{P}_\Omega^{Y|X} \odot \mathbb{P}_\Omega^{Z|YX}) \quad (72)$$

□

5.6 Conditional independence

Conditional independence has a familiar definition in probability models. It is sometimes possible to infer the existence of a uniform conditional probability from a conditional independence statement. Conditional independence can be equivalently defined either in terms of a factorisation of a joint probability distribution (Definition 5.25) or in terms of the existence of a conditional distribution that ignores one of its inputs (Theorem 5.26).

The latter formulation allows us, in some cases, to conclude from a the combination of a uniform conditional probability and a conditional independence statement the existence of a further uniform conditional probability (Corollary 5.28). We will discuss in Section 4 how uniform conditional probabilities can be thought of as causal relationships. Thus this means: from a fundamental assumed causal relationship and a conditional independence observed under the right conditions, we can conclude the existence of an additional causal relationship.

Definition 5.25 (Conditional independence). For a *probability model* \mathbb{P}_α and variables A, B, Z , we say B is conditionally independent of A given C , written $B \perp\!\!\!\perp_{\mathbb{P}_\alpha} A|C$, if

$$\mathbb{P}_\alpha^{ABC} = \begin{array}{c} \triangleleft \mathbb{P}_\alpha^C \quad \begin{array}{l} \boxed{\mathbb{P}_\alpha^{A|C}} \text{---} A \\ \boxed{\mathbb{P}_\alpha^{B|C}} \text{---} B \\ \text{---} C \end{array} \end{array} \quad (73)$$

Cho and Jacobs (2019) have shown that this definition coincides with the standard notion of conditional independence for a particular probability model (Theorem 5.26).

Conditional independence can equivalently be stated in terms of the existence of a conditional probability that “ignores” one of its inputs.

Theorem 5.26. *Given standard measurable (Ω, \mathcal{F}) , a probability model \mathbb{P} and variables $W : \Omega \rightarrow W$, $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, $Y \perp\!\!\!\perp_{\mathbb{P}} X|W$ if and only if there exists some version of $\mathbb{P}^{Y|WX}$ and $\mathbb{K} : W \rightarrow Y$ such that*

$$\mathbb{P}^{Y|WX} = \begin{array}{c} W \text{---} \boxed{\mathbb{K}} \text{---} Y \\ X \text{---} * \end{array} \quad (74)$$

$$\iff \mathbb{P}^{Y|WX}(A|w, x) = \mathbb{K}(A|w) \quad \forall A \in \mathcal{Y} \quad (75)$$

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

Theorem 5.27. *Given standard measurable (Ω, \mathcal{F}) , variables $W : \Omega \rightarrow W$, $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ and a probability set \mathbb{P}_C with uniform conditional probability $\mathbb{P}_C^{Y|WX}$ and $\alpha \in C$ such that $\mathbb{P}_\alpha^{WX} \gg \{\mathbb{P}_\beta^{WX} | \beta \in C\}$, $Y \perp\!\!\!\perp_{\mathbb{P}_\alpha} X|W$ if and only if there is a version of $\mathbb{P}_C^{Y|WX}$ and $\mathbb{K} : W \rightarrow Y$ such that*

$$\mathbb{P}_C^{Y|WX} = \begin{array}{c} W \text{---} \boxed{\mathbb{K}} \text{---} Y \\ X \text{---} * \end{array} \quad (76)$$

Proof. See Appendix \square

Corollary 5.28. *Given standard measurable (Ω, \mathcal{F}) , variables $W : \Omega \rightarrow W$, $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ and a probability set \mathbb{P}_C with uniform conditional probability $\mathbb{P}_C^{Y|WX}$ and $\alpha \in C$ such that $\mathbb{P}_\alpha^{WX} \gg \{\mathbb{P}_\beta^{WX} | \beta \in C\}$, $\mathbb{P}_C^{Y|W}$ exists if $Y \perp\!\!\!\perp_{\mathbb{P}_\alpha} X|W$.*

Proof. By Theorem 5.27, there is $\mathbb{K} : W \rightarrow Y$ such that for all β

$$\mathbb{P}_\beta^{WY} = \begin{array}{c} \begin{array}{c} \text{Diagram (77): A triangle labeled } \mathbb{P}_\alpha^{WX} \text{ has two output nodes. The top node connects to } W \text{ and } Y. \text{ The bottom node connects to } Y \text{ and } Y. \text{ A box labeled } \mathbb{P}_C^{Y|WX} \text{ has input } W \text{ and output } Y. \end{array} \\ (77) \end{array}$$

$$= \begin{array}{c} \text{Diagram (78): A triangle labeled } \mathbb{P}_\alpha^{WX} \text{ has two output nodes. The top node connects to } W \text{ and } Y. \text{ The bottom node connects to } Y \text{ and } Y. \text{ A box labeled } \mathbb{K} \text{ has input } W \text{ and output } Y. \end{array} \quad (78)$$

$$= \begin{array}{c} \text{Diagram (79): A triangle labeled } \mathbb{P}_\alpha^{W} \text{ has one output node connecting to } W \text{ and } Y. \text{ A box labeled } \mathbb{K} \text{ has input } W \text{ and output } Y. \end{array} \quad (79)$$

Thus \mathbb{K} is a version of $\mathbb{P}_C^{Y|W}$. \square

5.7 Uniform conditional independence

There are different notions of conditional independence that could be applied to a probability set \mathbb{P}_C . We can say X is “globally independent” of Y given Z if for every $\mathbb{P}_\alpha \in \mathbb{P}_C$, $X \perp\!\!\!\perp_{\mathbb{P}_\alpha} Y|Z$. Alternatively, we can say X is “uniformly independent” of Y given Z if $\mathbb{P}_C^{X|YZ}$ exists and does not depend on Y . We are particularly interested in the second kind, as this is the kind of conditional independence that enables simplified representations of uniform conditional distributions.

Both of these kinds of conditional independence are special cases of *extended conditional independence*, introduced by Constantinou and Dawid (2017). Extended conditional independence is a generalisation of conditional independence that is applicable to probability sets. In full generality, extended conditional independence makes use of the notion of “nonstochastic variables”, which are analogous to our notion of observed variables but applied to the set of choices C .

Extended conditional independence provides a unified way to express global conditional independence, uniform conditional independence and forms of conditional independence intermediate between the two. However, we only make use of uniform conditional independence in this work.

Definition 5.29 (Uniform conditional independence). Given a probability set \mathbb{P}_C and variables X , Y and Z , the uniform conditional independence $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e X|Z$

$XC|Z$ holds if $\mathbb{P}_C^{Y|XZ}$ and $\mathbb{P}_C^{Y|X}$ exist and

$$\begin{array}{ccc} & Z & \text{---} \boxed{\mathbb{P}_C^{Y|Z}} \text{---} Y \\ \mathbb{P}_C^{Y|XZ} \stackrel{\mathbb{P}_C}{\cong} & X & \text{---} * \end{array} \quad (80)$$

$$\iff \quad (81)$$

$$\mathbb{P}_C^{Y|XZ}(A|x, z) \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y|Z}(A|z) \quad \forall A \in \mathcal{Y}, (x, z) \in X \times Z \quad (82)$$

The notation $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e XC|Z$ is intentionally similar to a statement of extended conditional independence as defined by Constantinou and Dawid (2017). However, uniform conditional independence is a stronger assumption than extended conditional independence as the latter allows for arbitrary functions satisfy an equation like Eq. 82, while we require that these functions are Markov kernels (they are measurable and probability distribution-value, as in Definition 5.3).

Example 5.30 (Choice variable). Suppose we have a decision procedure $\mathcal{S}_C := \{\mathcal{S}_\alpha | \alpha \in C\}$ that consists of a measurement procedure for each element of a denumerable set of choices C . Each measurement procedure \mathcal{S}_α is modeled by a probability distribution \mathbb{P}_α on a shared sample space (Ω, \mathcal{F}) such that we have an observable “choice” variable $(D, D \circ \mathcal{S}_\alpha)$ where $D \circ \mathcal{S}_\alpha$ always yields α .

Furthermore, Define $Y : \Omega \rightarrow \Omega$ as the identity function. Then, by supposition, for each $\alpha \in A$, \mathbb{P}_α^{YC} exists and for $A \in \mathcal{Y}$, $B \in \mathcal{C}$:

$$\mathbb{P}_\alpha^{YC}(A \times B) = \mathbb{P}_\alpha(A)\delta_\alpha(B) \quad (83)$$

This implies, for all $\alpha \in C$

$$\mathbb{P}_\alpha^{Y|D} = \mathbb{P}_\alpha^Y \quad (84)$$

Thus $\mathbb{P}_C^{Y|D}$ exists and

$$\mathbb{P}_C^{Y|D}(A|\alpha) = \mathbb{P}_\alpha^Y(A) \quad \forall A \in \mathcal{Y}, \alpha \in C \quad (85)$$

Because only deterministic marginals \mathbb{P}_α^D are available, for every $\alpha \in C$ we have $Y \perp\!\!\!\perp_{\mathbb{P}_\alpha} D$. This reflects the fact that *after we have selected a choice* α the value of C provides no further information about the distribution of Y , because D is deterministic given any α . It does not reflect the fact that “choosing different values of C has no effect on Y ”.

Theorem 5.31 (Uniform conditional independence representation). *Given a*

probability set \mathbb{P}_C with a uniform conditional probability $\mathbb{P}_C^{\mathbf{XY}|Z}$,

$$\mathbb{P}_C^{\mathbf{XY}|Z} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{\mathbf{Y}|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{\mathbf{X}|Z}} \text{ --- } X \end{cases} \end{array} \quad (86)$$

$$\iff \quad (87)$$

$$\mathbb{P}_C^{\mathbf{XY}|Z}(A \times B|z) \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{\mathbf{X}|Z}(A|z)\mathbb{P}_C^{\mathbf{Y}|Z}(B|z) \quad \forall A \in \mathcal{X}, B \in \mathcal{Y}, z \in Z \quad (88)$$

if and only if $Y \perp_{\mathbb{P}_C}^e XC|Z$

Proof. If: By Theorem 8.4

$$\mathbb{P}_C^{\mathbf{XY}|Z} = \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{\mathbf{Y}|ZX}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{\mathbf{X}|Z}} \text{ --- } X \end{cases} \end{array} \quad (89)$$

$$\stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{\mathbf{Y}|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{\mathbf{X}|Z}} \text{ --- } X \end{cases} \end{array} \quad (90)$$

$$= \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{\mathbf{Y}|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{\mathbf{X}|Z}} \text{ --- } X \end{cases} \end{array} \quad (91)$$

Only if: Suppose

$$\mathbb{P}_C^{\mathbf{XY}|Z} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{\mathbf{Y}|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{\mathbf{X}|Z}} \text{ --- } X \end{cases} \end{array} \quad (92)$$

and suppose for some $\alpha \in C$, $A \times C \in \mathcal{X} \otimes \mathcal{Z}$, $B \in \mathcal{Y}$ $\mathbb{P}_\alpha^{\mathbf{XZ}}(A \times C) > 0$ and

$$\mathbb{P}_C^{\mathbf{Y|XZ}}(B|x, z) > \mathbb{P}_C^{\mathbf{Y|Z}}(B|z) \quad \forall (x, z) \in A \times C \quad (93)$$

then

$$\mathbb{P}_\alpha^{\mathbf{XYZZ}}(A \times B \times C) = \int_{A \times C} \mathbb{P}_C^{\mathbf{Y|XZ}}(B|x, z) \mathbb{P}_C^{\mathbf{X|Z}}(dx|z) \mathbb{P}_\alpha^{\mathbf{Z}}(dz) \quad (94)$$

$$> \int_{A \times C} \mathbb{P}_C^{\mathbf{Y|X}}(B|z) \mathbb{P}_C^{\mathbf{X|Z}}(dx|z) \mathbb{P}_\alpha^{\mathbf{Z}}(dz) \quad (95)$$

$$= \int_C \mathbb{P}_C^{\mathbf{XY|X}}(A \times B|z) \mathbb{P}_\alpha^{\mathbf{Z}}(dz) \quad (96)$$

$$= \mathbb{P}_\alpha^{\mathbf{XYZZ}}(A \times B \times C) \quad (97)$$

a contradiction. An analogous argument follows if we replace “>” with “<” in Eq. 93. \square

6 When do response conditionals exist?

Lemmas are intermediate steps

Our approach is to model decision problems with probability sets \mathbb{P}_C for some set of choices C . If we have a pair of variables X and Y such that $\mathbb{P}_C^{Y|X}$ exists, then the model says that the joint outcome \mathbb{P}_α^{XY} of any choice $\alpha \in C$ can be computed from the marginal distribution \mathbb{P}_α^X alone. We are going to ask the question: in which kind of probability sets do uniform conditionals of the form $\mathbb{P}_C^{Y|XH}$ exist? Here H is a “fixed but unknown” hypothesis that becomes better known as more data is observed. Roughly speaking, $\mathbb{P}_C^{Y|XH}$ represents the response of Y to X regardless of which choice is made.

A decision makers may be interested in a functions like $\mathbb{P}_C^{Y|XH}$. Suppose they have substantial prior knowledge about how to control X , less knowledge about controlling Y and access to a sequence of data points. If the data points can identify H , then If a decision maker has prior knowledge of how to control X and data that is informative about the value of H , these pieces of knowledge together can help them to control Y . We call a uniform conditional of the form $\mathbb{P}_C^{Y|XH}$ a *response conditional*.

If a model \mathbb{P}_C supports $\mathbb{P}_C^{Y|XH}$, it also provides an answer to the concerns of Hernán and Taubman. Paraphrasing their argument without the use of potential outcomes: they were concerned that there may be different ways to achieve particular values or distributions over X , and these may also lead to different distributions over Y . In that case, they argued that there was no well-defined causal effect of X on Y . However, if $\mathbb{P}_C^{Y|XH}$ then the model describes a situation where – once we have enough data to pin down H – it doesn’t matter any more which particular choice leads to a given distribution over X .

We consider first the question of what kind of probability sets \mathbb{P}_C support uniform conditional distributions of the form $\mathbb{P}_C^{Y|XH}$. Secondly, we consider what kind of decision procedures can be modeled by probability sets of this type.

6.1 Sequential decision models

In order to pose this question, we need a setting in which we expect to observe sequential data. That is, our model is a probability set \mathbb{P}_C on sample space (Ω, \mathcal{F}) such that we have variables $Y := (Y_i)_{i \in M}$ (the outcome sequence) and $D := (D_i)_{i \in M}$ (the action sequence) for some index set $M \subset \mathbb{N}$. We say Y_i corresponds to D_i . We are specifically looking for uniform conditionals of the form $\mathbb{P}_C^{Y_i|D_iH}$ for all $i \in M$. H here is a hypothesis, and it must be fixed with respect to different choices – i.e. $H \perp\!\!\!\perp_{\mathbb{P}_C^e} C$.

We assume a starting point that $\mathbb{P}_C^{Y|D}$ exists. This could be guaranteed, for example, if each choice α corresponds to a unique deterministic distribution \mathbb{P}_α^D (see Example 5.30).

There are two further assumptions relevant to the existence of response conditionals. The first is *exchange commutativity*. This is the condition that we

get the same result from applying a swap transformation to the input of $\mathbb{P}_C^{Y|D}$ as we get from applying the same swap transformation to its output.

The second is a condition of *consequence locality*. This is the assumption that, for any $A \subset M$, $\mathbb{P}_C^{Y_A|D_A}$ exists and

$$\mathbb{P}_C^{Y_A|D_M} = \begin{array}{c} D_A \text{ --- } \boxed{\mathbb{P}_C^{Y_A|D_A}} \text{ --- } Y_A \\ D_{M \setminus A} \text{ --- } * \end{array} \quad (98)$$

In the language of extended conditional independence, it is the assumption $Y_A \perp\!\!\!\perp_{\mathbb{P}_C}^e CD_{M \setminus A} | D_A$.

Exchange commutativity is similar, but not identical, to a number of assumptions discussed in the literature. *Post-treatment exchangeability* found in Dawid (2021) is implied by exchange commutativity, but not the reverse. There are also notions of “causal exchangeability” found in Greenland and Robins (1986) and Banerjee et al. (2017); a subtle difference between these notions and exchange commutativity is that these latter notions are symmetries of *procedures* – they involve actually swapping actions or individuals in an experiment – while exchange commutativity is a symmetry of a probability set.

Consequence locality is similar to the stable unit treatment distribution assumption (SUTDA) in Dawid (2021). 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 sub-assumptions. 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.

Both SUTDA and SUTVA talk about how an outcome Y_i does not depend on, or is not affected by, any of the actions that do not correspond to it. Such statements would need to be made more precisely if we want to evaluate what precise relation they have to consequence locality.

Put the following in the discussion of decision procedures

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 similar to 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.

Put the preceding in the discussion of decision procedures

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

The difference between exchangeability (de Finetti, [1937] 1992) and exchange commutativity is illustrated by the following pair of diagrams. Exchangeability is a symmetry of probability distributions – a distribution is exchangeable if it is unchanged by swapping outputs. Exchange commutativity is a symmetry of Markov kernels – a Markov kernel is exchange commutative if swapping inputs and swapping outputs gives the same result.

Exchangeability (swapping labels):

$$(99)$$

Exchange commutativity (swapping choices \sim swapping labels):

$$(100)$$

—end not sure where to put—

6.2 Causal contractibility

Here we set out formal definitions of exchange commutativity and locality of consequences, as well as “consequence contractibility”, which is the conjunction of both conditions.

Definition 6.1 (Swap map). Given $M \subset \mathbb{N}$ a finite permutation $\rho : M \rightarrow M$ and a variable $\mathbf{X} : \Omega \rightarrow X^M$ such that $\mathbf{X} = (\mathbf{X}_i)_{i \in M}$, define the Markov kernel $\text{swap}_{\rho(\mathbf{X})} : X^M \rightarrow X^M$ by $(d_i)_{i \in \mathbb{N}} \mapsto \delta_{(d_{\rho(i)})_{i \in \mathbb{N}}}$.

Definition 6.2 (Exchange commutativity). Suppose we have a sample space (Ω, \mathcal{F}) and a probability set \mathbb{P}_C with uniform conditional probability $\mathbb{P}_C^{\mathbf{Y} | \mathbf{D}}$ where $\mathbf{Y} := \mathbf{Y} := (\mathbf{Y}_i)_M$, $\mathbf{D} := \mathbf{D}_M := (\mathbf{D}_i)_M$, $M \subseteq \mathbb{N}$. If for any finite permutation $\rho : M \rightarrow M$

$$\text{swap}_{\rho(\mathbf{D})} \mathbb{P}_C^{\mathbf{Y} | \mathbf{D}} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{\mathbf{Y} | \mathbf{D}} \text{swap}_{\rho(\mathbf{Y})} \quad (101)$$

Then $\mathbb{P}_C^{\mathbf{Y} | \mathbf{D}}$ commutes with exchange.

If $\mathbb{P}_C^{Y|D}$ commutes with exchange and we have $\alpha, \alpha' \in C$ such that $\mathbb{P}_\alpha^C = \mathbb{P}_{\alpha', \text{swap}_{\rho(D)}}^C$, then $\mathbb{P}_\alpha^Y = \mathbb{P}_{\alpha', \text{swap}_{\rho(Y)}}^Y$. However, $\mathbb{P}_C^{Y|D}$ may commute with exchange even if there are no such α and $\alpha' \in C$.

Definition 6.3 (Locality of consequences). Suppose we have a sample space (Ω, \mathcal{F}) and a probability set \mathbb{P}_C with uniform conditional probability $\mathbb{P}_C^{Y|D}$ where $Y := Y := (Y_i)_M$, $D := D_M := (D_i)_M$, $M \subseteq \mathbb{N}$. If for any $A \subset M$

$$\mathbb{P}_S^{Y_A|D_M} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} D_A \text{ --- } \boxed{\mathbb{P}_C^{Y_A|D_A}} \text{ --- } Y_A \\ D_{M \setminus A} \text{ --- } * \end{array} \quad (102)$$

then $\mathbb{P}_C^{Y|D}$ exhibits *consequence locality*.

If $\mathbb{P}_C^{Y|D}$ exhibits consequence locality then, given two different choices α and α' such that $\mathbb{P}_\alpha^{D_A} = \mathbb{P}_{\alpha'}^{D_A}$ then $\mathbb{P}_\alpha^{Y_A} = \mathbb{P}_{\alpha'}^{Y_A}$. However, once again, $\mathbb{P}_C^{Y|D}$ may exhibit consequence locality even if no such pair of choices exists.

Neither condition implies the other.

Theorem 6.4. *Exchange commutativity does not imply locality of consequences or vice 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 probability set \mathbb{P}_C with uniform conditional $\mathbb{P}_C^{Y|D}$, where $D = (D_1, D_2)$, $Y = (Y_1, Y_2)$.

Suppose the unique version of $\mathbb{P}_C^{Y|D}$ is

$$\mathbb{P}_C^{Y|D}(y_1, y_2 | d_1, d_2) = \mathbb{I}[(y_1, y_2) = (d_1 + d_2, d_1 + d_2)] \quad (103)$$

then

$$\mathbb{P}_C^{Y_1|D}(y_1 | d_1, d_2) = \mathbb{I}[y_1 = d_1 + d_2] \quad (104)$$

and there is no function depending on y_1 and d_1 only that is equal to this. Thus \mathbb{P}_C exhibits non-local consequences.

However, taking ρ to be the unique nontrivial swap $\{0, 1\} \rightarrow \{0, 1\}$

$$\text{swap}_{\rho(D)} \mathbb{P}_C^{Y|D}(y_1, y_2 | d_1, d_2) = \mathbb{P}_C^{Y|D}(y_1, y_2 | d_2, d_1) \quad (105)$$

$$= \mathbb{I}[(y_1, y_2) = (d_2 + d_1, d_2 + d_1)] \quad (106)$$

$$= \mathbb{I}[(y_1, y_2) = (d_1 + d_2, d_1 + d_2)] \quad (107)$$

$$= \mathbb{I}[(y_2, y_1) = (d_1 + d_2, d_1 + d_2)] \quad (108)$$

$$= \mathbb{P}_C^{Y|D} \text{swap}_{\rho(Y)}(y_1, y_2 | d_1, d_2) \quad (109)$$

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

A conditional probability model that exhibits locality of consequences but does not commute with exchange follows. Suppose again $D = Y = \{0, 1\}$ and we have a probability set \mathbb{P}_C with uniform conditional $\mathbb{P}_C^{Y|D}$, where $D = (D_1, D_2)$, $Y = (Y_1, Y_2)$. This time, suppose the unique version of $\mathbb{P}_C^{Y|D}$ is

$$\mathbb{P}_C^{Y|D}(y_1, y_2 | d_1, d_2) = \mathbb{I}[(y_1, y_2) = (0, 1)] \quad (110)$$

Then If $\mathbb{P}_\alpha^{D^S} = \mathbb{P}_\beta^{D^S}$ for $S \subset \{0, 1\}$ then:

$$\mathbb{P}_C^{Y_1|D}(y_1 | d_1, d_2) = \mathbb{I}[y_1 = 0] \quad (111)$$

$$= \mathbb{P}_C^{Y_1|D_1}(y_1 | d_1) \quad (112)$$

$$\mathbb{P}_C^{Y_2|D}(y_2 | d_1, d_2) = \mathbb{I}[y_2 = 1] \quad (113)$$

$$= \mathbb{P}_C^{Y_2|D_2}(y_2 | d_2) \quad (114)$$

so $\mathbb{P}_C^{Y|D}$ exhibits consequence locality.

However, \mathbb{P}_C does not commute with exchange.

$$\text{swap}_{\rho(D)} \mathbb{P}_C^{Y|D}(y_1, y_2 | d_1, d_2) = \mathbb{P}_C^{Y|D}(y_1, y_2 | d_2, d_1) \quad (115)$$

$$= \mathbb{I}[(y_1, y_2) = (0, 1)] \quad (116)$$

$$\neq \mathbb{I}[(y_2, y_1) = (0, 1)] \quad (117)$$

$$= \mathbb{P}_C^{Y|D} \text{swap}_{\rho(D)}(y_1, y_2 | d_1, d_2) \quad (118)$$

□

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 $D_1 = 0$), 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 $D_1 = 1$). At the second step, I may further choose to copy the provisional results ($D_2 = 0$) or invert them ($D_2 = 1$). Then

$$\mathbb{P}_S^{Y_1|D}(y_1 | d_1, d_2) = 0.5 \quad (119)$$

$$\mathbb{P}_S^{Y_2|D}(y_2 | d_1, d_2) = 0.5 \quad (120)$$

- 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 satisfy Definition 6.3
- 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 locality *causal contractibility*.

Definition 6.5 (Causal contractibility). A probability set \mathbb{P}_S with uniform conditional $\mathbb{P}_C^{Y|D}$ is $(D; Y)$ causally contractible if it is both exchange commutative and exhibits consequence locality.

Theorem 6.6 (Causal contractibility). A probability set \mathbb{P}_S with uniform conditional $\mathbb{P}_C^{Y|D}$ is (D, Y) -causally contractible if and only if for any $A, B \subset M$ with $|A| = |B|$

$$\mathbb{P}_C^{Y_A|D_A D_{M \setminus A}} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_B|D_B D_{M \setminus B}} \quad (121)$$

$$\begin{array}{ccc} D_A & \xrightarrow{\quad \mathbb{P}_C^{Y_A|D_A} \quad} & Y_A \\ \mathbb{P}_C \cong \downarrow & & \\ D_{M \setminus A} & \xrightarrow{\quad * \quad} & \end{array} \quad (122)$$

Proof. If: Choosing $A = B$ yields consequence locality immediately from Eq. 122.

Also from Eq. 122, we have

$$\mathbb{P}_C^{Y_A|D_A} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_B|D_B} \quad (123)$$

For any permutation $\rho : M \rightarrow M$, take A such that $|A| = |M|$ and $B = \rho(A)$. Then

$$\mathbb{P}_C^{Y_A|D_A} \stackrel{\mathbb{P}_C}{\cong} \text{swap}_{\rho^{-1}(D)} \mathbb{P}_C^{Y_A|D_A} \text{swap}_{\rho(Y)} \quad (124)$$

So \mathbb{P}_C commutes with exchange.

Only if: For any $A, B \subset M$, let $s_{BA} : D^M \rightarrow D^M$ be the swap map that sends the B indices to A indices and $s_{AB} : Y^M \rightarrow Y^M$ be the swap map that sends A indices to B indices.

$$\begin{array}{ccc} D_A & \xrightarrow{\quad \mathbb{P}_C^{Y_A|D_A} \quad} & Y_A \\ D_{M \setminus A} & \xrightarrow{\quad * \quad} & \end{array} = \begin{array}{ccc} D_A & \xrightarrow{\quad \mathbb{P}_C^{Y_A Y_{M \setminus A} | D_A D_{M \setminus A}} \quad} & Y_A \\ D_{M \setminus A} & \xrightarrow{\quad * \quad} & \end{array} \quad (125)$$

$$= \begin{array}{ccc} D_A & \xrightarrow{\quad s_{BA} \quad} & \mathbb{P}_C^{Y_A Y_{M \setminus A} | D_A D_{M \setminus A}} \xrightarrow{\quad s_{AB} \quad} Y_A \\ D_{M \setminus A} & \xrightarrow{\quad * \quad} & \end{array} \quad (126)$$

$$= \begin{array}{ccc} D_B & \xrightarrow{\quad \mathbb{P}_C^{Y_B Y_{M \setminus B} | D_B D_{M \setminus B}} \quad} & Y_B \\ D_{M \setminus B} & \xrightarrow{\quad * \quad} & \end{array} \quad (127)$$

□

Corollary 6.7. *A causally contractible probability set \mathbb{P}_S with uniform conditional $\mathbb{P}_C^{Y|D}$ has the property*

$$\mathbb{P}_C^{Y_A|D_A} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_B|D_B} \quad (128)$$

for all $A, B \subset M$.

I think this condition doesn't imply consequence locality, but haven't thought of a counterexample yet

6.3 Existence of response conditionals

The main result in this section is Theorem 6.10 which shows that a probability set \mathbb{P}_C is causally contractible if and only if it can be represented as the product of a distribution over hypotheses \mathbb{P}_\square^H and a collection of identical uniform conditionals $\mathbb{P}_C^{Y_1|D_1^H}$. 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” $Y_1|D_1$ conditional probability.

Lemma 6.8. *Given a probability set $\mathbb{P}_C^{Y|D}$ such that $D := (D_i)_{i \in \mathbb{N}}$ and $Y := (Y_i)_{i \in \mathbb{N}}$, \mathbb{P}_\square is consequence contractible if and only if there exists a column exchangeable probability distribution $\mu^{Y^D} \in \Delta(Y^{D \times \mathbb{N}})$ such that*

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \triangle \\ \mu^{Y^D} \\ D \end{array} \xrightarrow{\quad} \boxed{\mathbb{F}_{ev}} \xrightarrow{\quad} Y \quad (129)$$

$$\iff \quad (130)$$

$$\mathbb{P}_C^{Y|D}(y|(d_i)_{i \in \mathbb{N}}) = \mu^{Y^D} \Pi_{(d_i)_{i \in \mathbb{N}}}(y) \quad (131)$$

Where $\Pi_{(d_i)_{i \in \mathbb{N}}} : Y^{D \times \mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is the function that projects the (d_i, i) indices for all $i \in \mathbb{N}$ and \mathbb{F}_{ev} is the Markov kernel associated with the evaluation map

$$ev : D^{\mathbb{N}} \times Y^{D \times \mathbb{N}} \rightarrow Y \quad (132)$$

$$((d_i)_{i \in \mathbb{N}}, (y_{ij})_{i \in D, j \in \mathbb{N}}) \mapsto (y_{d_i i})_{i \in \mathbb{N}} \quad (133)$$

Proof. Only if: Consider a probability set $\mathbb{P}_{C'}$ where $C' \supset C$ contains all α such that \mathbb{P}_α^D is deterministic and $\mathbb{P}_{C'}^{Y|D} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y|D}$. It can be constructed by adding to \mathbb{P}_C probability sets with marginals $\delta_d \odot \mathbb{P}_{C'}^{Y|D}$ for all $d \in D$.

We will prove the result holds for $\mathbb{P}_{C'}$, and it will therefore also hold for \mathbb{P}_C .

For all $d \in D$, abuse notation to say that \mathbb{P}_d is a probability set in C' such that $\mathbb{P}_d^D = \delta_d$. For any $\alpha \in C'$, we have

$$\mathbb{P}_\alpha^{\text{DY}}(B \times C) = \int_B \mathbb{P}_C^{\text{Y|D}}(C|d) \mathbb{P}_\alpha^{\text{D}}(dd) \quad (134)$$

$$= \int_B \int_D \mathbb{P}_C^{\mathbf{Y}|D}(C|d') \mathbb{P}_d^D(dd') \mathbb{P}_\alpha^D(dd) \quad (135)$$

$$= \int_B \mathbb{P}_d^Y(C) \mathbb{P}_\alpha^D(dd) \quad (136)$$

Thus $d \mapsto \mathbb{P}_d^Y$ is a version of $\mathbb{P}_C^{Y|C}$.

Choose $e := (e_i)_{i \in \mathbb{N}}$ such that $e_{|D|+j}$ is the i th element of D for all $i, j \in \mathbb{N}$. Define

$$\mu^{Y^D}((y_{ij})_{D \times \mathbb{N}}) := \mathbb{P}_e^Y((y_{|D|i+j})_{i \in D, j \in \mathbb{N}}) \quad (137)$$

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

Define

$$\mathbb{Q} : D \rightarrowtail Y \quad (138)$$

$$Q := \begin{array}{c} \triangle \\ \mu^{Y^D} \\ \text{D} \end{array} \xrightarrow{\quad} \boxed{\text{F}_{\text{ev}}} \xrightarrow{\quad} Y \quad (139)$$

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

$$\sum_{y \in Y^{-1}(y_A)} \mathbb{Q}(y|d) = \sum_{y \in Y^{-1}(y_A)} \mu^{(Y_{d_{ii}}^D)^A}(y) \quad (140)$$

$$= \sum_{y \in Y^{-1}(y_A)} \mathbb{P}_e^{(Y|D|d_i+i)A}(y) \quad (141)$$

$$= \mathbb{P}_e^{\mathbf{Y}_B}(y_A) \quad (142)$$

$$= \mathbb{P}_d^{\mathbf{Y}^A}(y_A) \quad \text{by causal contractibility} \quad (143)$$

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

$$\mathbb{Q}(y|d) = \mathbb{P}_d^{\mathbf{Y}}(y) \quad (144)$$

And so \mathbb{Q} is also a version of $\mathbb{P}_{\square}^{\mathbf{Y}|\mathbf{C}}$.

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

$$\mu^{Y^D} \Pi_S = \mathbb{P}_e^{Y_{\rho(S)}} \quad (145)$$

$$= \mathbb{P}_e^{Y_{\rho(T)}} \quad \text{by contractibility of } \mathbb{P}_C \text{ and the equality } e_{\rho(S)} = e_{\rho(T)} \quad (146)$$

$$= \mu^{Y^D} \Pi_T \quad (147)$$

If: Suppose

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \triangle \\ \mu^{Y^D} \\ \text{D} \end{array} \begin{array}{c} \text{---} \\ \text{F}_{ev} \end{array} \text{---} Y \quad (148)$$

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

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

By definition,

$$\mathbb{P}_C^{Y_S|D}(y_S|d) = \sum_{y_S^D \in Y^{|D| \times |S|}} \mu^{Y^D} \Pi_S(y_S^D) \mathbb{F}_{ev}(y_S|d, y_S^D) \quad (149)$$

$$= \sum_{y_S^D \in Y^{|D| \times |T|}} \mathbb{P}_C^{Y_T^D}(y_S^D) \mathbb{F}_{ev}(y_S|d, y_S^D) \quad \text{by contractibility of } \mu^{Y^D} \Pi_T \quad (150)$$

$$= \mathbb{P}_C^{Y_T|D}(y_S|d) \quad (151)$$

□

It is useful to apply conventions established for discussing variables to the tabular distribution μ^{Y^D} and the representation of $\mathbb{P}_C^{Y|D}$ in Equation 129. Thus we define an augmented causally contractible model as follows:

Definition 6.9 (Augmented causally contractible model). A (D, Y) -causally contractible probability set \mathbb{P}_C on (Ω, \mathcal{F}) is *augmented* if there is an unobserved variable $Y^D : \Omega \rightarrow Y^{|\mathbb{N}| \times |D|}$ such that $\mathbb{P}_C^{Y^D}$ exists and

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \triangle \\ \mathbb{P}_C^{Y^D} \\ \text{D} \end{array} \begin{array}{c} \text{---} \\ \text{F}_{ev} \end{array} \text{---} Y \quad (152)$$

An augmented causally contractible model looks in some respects similar to a potential outcomes model - both have a distribution over an unobserved tabular variable (Y^D or the potential outcomes respectively), and the value of Y_i is deterministically equal to the entry in the table corresponding to (i, D) .

However, the Y^D in an augmented causally contractible model usually can't be interpreted as potential outcomes. For example, consider a series of bets on fair coin flips. Model the consequence Y_i as uniform on $\{0, 1\}$ for any decision D_i , for all i . Specifically, $D = Y = \{0, 1\}$ and $\mathbb{P}_\alpha^{Y^n}(y) = \prod_{i \in [n]} 0.5$ for all n , $y \in Y^n$, $\alpha \in R$. Then the construction of \mathbb{P}^{Y^D} following the method in Lemma 6.8 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 $Y_i^0 = 1 - Y_i^1$.

The following is the main theorem of this section, that establishes the equivalence between causal contractibility and the existence of response conditionals. The argument in outline is: because $\mathbb{P}_C^{Y^D}$ is a column exchangeable probability distribution we can apply De Finetti's theorem to show $\mathbb{P}_C^{Y^D}$ is representable as a product of identical parallel copies of $\mathbb{P}_C^{Y_1^D|H}$ and a common prior \mathbb{P}_C^H . This in turn can be used to show that $\mathbb{P}_C^{Y^D}$ can be represented as a product of identical parallel copies of $\mathbb{P}_C^{Y_i|D_i H}$ and the same common prior \mathbb{P}_C^H .

Theorem 6.10. *Suppose we have a sample space (Ω, \mathcal{F}) and a probability set \mathbb{P}_C with uniform conditional $\mathbb{P}_C^{Y^D}$ where $D := (D_i)_{i \in \mathbb{N}}$ and $Y := (Y_i)_{i \in \mathbb{N}}$. \mathbb{P}_C is augmented $(D; Y)$ -causally contractible if and only if there exists some $H : \Omega \rightarrow H$ such that \mathbb{P}_C^H and $\mathbb{P}_C^{Y_i|HD_i}$ exist for all $i \in \mathbb{N}$ and*

$$\mathbb{P}_C^{Y^D} = \begin{array}{c} \begin{array}{c} \triangle \mu^H \\ \text{---} \end{array} \begin{array}{c} \text{---} \\ \text{---} \end{array} \begin{array}{c} \boxed{\Pi_{D,i}} \quad \boxed{\mathbb{P}_{\square}^{Y_0|HD_0}} \\ \text{---} \end{array} Y_i \\ i \in \mathbb{N} \end{array} \quad (153)$$

$$\iff \quad (154)$$

$$Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e Y_{N \setminus i}, D_{N \setminus i} C | HD_i \quad \forall i \in \mathbb{N} \quad (155)$$

$$\wedge H \perp\!\!\!\perp_{\mathbb{P}_C}^e DC \quad (156)$$

$$\wedge \mathbb{P}_C^{Y_i|HD_i} = \mathbb{P}^{Y_0|HD_0} \quad \forall i \in \mathbb{N} \quad (157)$$

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

Proof. We make use of Lemma 6.8 to show that we can represent the conditional probability $\mathbb{P}_C^{Y^D}$ as

$$\mathbb{P}_C^{Y^D} = \begin{array}{c} \begin{array}{c} \triangle \mu^{Y^D} \\ \text{---} \end{array} \begin{array}{c} \text{---} \\ \text{---} \end{array} \boxed{\mathbb{F}_{ev}} \text{---} Y \\ D \end{array} \quad (158)$$

$$(159)$$

As a preliminary, we will show

$$\mathbb{F}_{\text{ev}} = D \quad \text{---} \quad \boxed{\begin{array}{c} \text{---} \Pi_{Y^D,i} \text{---} \\ \text{---} \Pi_{D,i} \text{---} \\ \text{---} \mathbb{F}_{\text{ev},i} \text{---} \end{array}} \quad Y_i \quad i \in \mathbb{N} \quad (160)$$

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

$$((y_i)_{i \in D}, d) \mapsto y_d \quad (161)$$

Recall that ev is the function

$$((d_i)_{i \in \mathbb{N}}, (y_{ij})_{i \in D, j \in \mathbb{N}}) \mapsto (y_{d_i i})_{i \in \mathbb{N}} \quad (162)$$

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

$$\mathbb{F}_{\text{ev}}\left(\prod_{i \in \mathbb{N}} A_i | (d_i)_{i \in \mathbb{N}}, (y_{ij})_{i \in D, j \in \mathbb{N}}\right) = \delta_{(y_{d_i i})_{i \in \mathbb{N}}} \left(\prod_{i \in \mathbb{N}} A_i\right) \quad (163)$$

$$= \prod_{i \in \mathbb{N}} \delta_{y_{d_i i}}(A_i) \quad (164)$$

$$= \text{copy}^{\mathbb{N}} \prod_{i \in \mathbb{N}} (\Pi_{D,i} \otimes \Pi_{Y,i}) \mathbb{F}_{\text{ev}_{Y^D \times D}} \quad (165)$$

Which is what we wanted to show.

Only if: As we have an augmented causally contractible model, we have a variable $Y^D = (Y_i^D)_{i \in \mathbb{N}}$ exchangeable with respect to $\mathbb{P}_C^{Y^D}$ (Lemma 6.8). From kal (2005) we have a directing random measure H such that

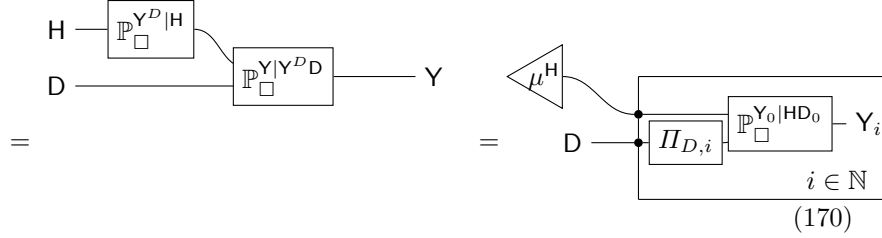
$$\mathbb{P}_C^{Y^D|H} = H \quad \text{---} \quad \boxed{\begin{array}{c} \text{---} \mathbb{P}_{\square}^{Y^D|H} \text{---} \\ \text{---} Y_i \end{array}} \quad i \in \mathbb{N} \quad (166)$$

$$\iff \quad (167)$$

$$\mathbb{P}_C^{Y^D|H} \left(\prod_{i \in \mathbb{N}} A_i | h \right) = \prod_{i \in \mathbb{N}} \mathbb{P}_C^{Y_i^D|H} (A_i | h) \quad (168)$$

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

$$\mathbb{P}_C^{Y|HD} = \mathbb{P}_C^{Y^D|HD} \odot \mathbb{P}_C^{Y|Y^DHD} \quad (169)$$



If: By assumption

$$\mathbb{P}_C^{Y|D}(\prod_{i \in \mathbb{N}} A_i | h, (d_i)_{i \in \mathbb{N}}) = \int_H \prod_{i \in \mathbb{N}} \mathbb{P}_C^{Y_1|HD_1}(A_i | h, d_i) \mathbb{P}_C^H(dh) \quad (171)$$

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

$$\mathbb{P}_\alpha^{Y_M}(A) = \int_{D^\mathbb{N}} \mathbb{P}_\alpha^{Y_M|D}(A|d) \mathbb{P}_\alpha^D(dd) \quad (172)$$

$$= \int_H \int_{D^\mathbb{N}} \prod_{i \in M} \mathbb{P}_C^{Y_1|HD_1}(A_i | h, d_i) \mathbb{P}_\alpha^D(dd) \mathbb{P}_C^H(dh) \quad (173)$$

$$= \int_H \int_{D^{|M|}} \prod_{i \in M} \mathbb{P}_C^{Y_1|HD_1}(A_i | h, d_i) \mathbb{P}_\alpha^{D_M}(dd_M) \mathbb{P}_C^H(dh) \quad (174)$$

$$= \int_H \int_{D^{|M|}} \prod_{i \in M} \mathbb{P}_C^{Y_1|HD_1}(A_i | h, d_i) \mathbb{P}_{\alpha'}^{D_N}(dd_N) \mathbb{P}_C^H(dh) \quad (175)$$

$$= \int_H \int_{D^\mathbb{N}} \prod_{i \in M} \mathbb{P}_C^{Y_1|HD_1}(A_i | h, d_i) \mathbb{P}_{\alpha'}^D(dd) \mathbb{P}_C^H(dh) \quad (176)$$

$$= \mathbb{P}_{\alpha'}^{Y_M}(A) \quad (177)$$

□

6.4 Modelling different decision procedures

We have a formal condition – causal contractibility – equivalent to the existence of response conditionals. However, a key challenge is to evaluate when a decision procedure is appropriately modeled with a probability set causally contractible with respect to some pair of variables. This is an opaque problem, and we don't expect that this can be avoided entirely.

Judgements of exchangeability (a more basic notion than causal contractibility) are justified on the basis of “symmetric ignorance” ?, chap. 5. More

specifically, if we are committed to modelling uncertainty with a single probability distribution and we cannot identify any knowledge that would lead us to construct a different model for:

- The original measurement procedure
- The original measurement procedure, composed with a function permuting the labels

then we should use an exchangeable probability distribution to model the original measurement procedure.

Causal contractibility judgements can appeal to similar considerations, but in practice there seem to be a greater number of subtle considerations that rule out judgements of causal contractibility.

Causal contractibility is a very strong assumption. Suppose we have a decision procedure in which M observations are made $(\mathcal{X}_M, \mathcal{Y}_M)$ that are unaffected by the choice α , followed by M repetitions $(\mathcal{X}_{(M,2M]}, \mathcal{Y}_{(M,2M]})$ which are responsive to the choice α . We model this with a probability set \mathbb{P}_C where \mathbf{X}_M corresponds to \mathcal{X}_M and so forth.

If an $(\mathbf{X}_{2M}, \mathbf{Y}_{2M})$ -causally contractible model \mathbb{P}_C is chosen, then the following holds (see corollary 6.7):

$$\mathbb{P}_C^{\mathbf{Y}_{[2,M+1]}|\mathbf{X}_{[2,M+1]}} = \mathbb{P}^{\mathbf{Y}_{(M,2M]}|\mathbf{X}_{(M,2M]}} \quad (178)$$

$$\implies \mathbb{P}_C^{\mathbf{Y}_{M+1}|\mathbf{X}_{[2,M+1]}\mathbf{Y}_{[2,M]}} = \mathbb{P}^{\mathbf{Y}_{M+1}|\mathbf{X}_{(M,2M]}\mathbf{Y}_{(M+1,2M]}} \quad (179)$$

That is, causal contractibility in this implies that there is no difference between conditioning on observational results or on the results of active choices with respect to predicting the result of a particular choice. We could not “reject the hypothesis of causal contractibility” by observing that one subsequence appears to differ from the other, as the model treats both subsequences identically.

We will reason about decision procedures in the following way:

- Propose a decision procedure \mathcal{S} and assume a conditional probability model \mathbb{P}_\square associated with it
- Consider an alternative decision procedure \mathcal{S}' and argue that a conditional probability model \mathbb{P}'_\square , related somehow to \mathbb{P}_\square , is appropriate to model it
- If, in addition, we accept that *the same model* is appropriate for both \mathcal{S} and \mathcal{S}' , then we have $\mathbb{P}_\square = \mathbb{P}'_\square$

An example of this reasoning is found in discussions of exchangeability, outside the setting of decision problems. Suppose we have a measurement procedure \mathcal{S} and an observed variable $(\mathbf{Y} \circ \mathcal{S}, \mathbf{Y})$ where $\mathbf{Y} = (\mathbf{Y}_i)_{i \in \mathbb{N}}$ for $\mathbf{Y}_i : \Omega \rightarrow Y$ – i.e. \mathcal{S} yields a countably infinite sequence of values from the set Y . Consider an alternative observed variable $(\mathbf{Y}' \circ \mathcal{S}, \mathbf{Y}')$ where $\mathbf{Y}' = \text{swap} \circ \mathbf{Y}$. \mathcal{Y}' is the procedure “do whatever is done to measure \mathcal{Y} , then shuffle the results according to swap”.

\mathcal{S} is modeled by some ordinary probability model $(\mu, \Omega, \mathcal{F})$, with marginal distribution μ^Y of Y . It may be the case that we think that the additional shuffle operation involved in \mathcal{Y}' does not in any way alter the appropriate probability model associated with the variable. Thus $\mu^Y = \mu^{Y'}$, and the model μ is exchangeable with respect to Y .

In the following section, we will apply this approach to considering whether the assumption of exchange commutativity is justified for a given decision procedure. The approach outlined can be used to assess whether measurement procedures support exchange commutativity or consequence locality. However, we will focus on exchange commutativity because, out of the two assumptions, we have more to say about it.

6.5 Decision procedures with response conditionals

Suppose we have a decision procedure \mathcal{S}_A . We consider an exchange commutative model appropriate if we think that the experiment given by

- Suppose α has been decided on
- Obtain a sequence of choices according to \mathcal{D} (which depends on α) and enact them
- Measure the consequences according to \mathcal{Y}

is equivalent to \mathcal{S}' (in the sense of requiring the same model):

- Decide on α' such that $\mathbb{P}_{\alpha'}^D = \mathbb{P}_{\alpha}^D$
- Obtain a sequence of choices according to \mathcal{D} (which depends on α') and enact them
- Measure preliminary consequences according to \mathcal{Y}
- Apply the inverse shuffle swap^{-1} to the preliminary consequences to get the consequences of interest \mathcal{Y}'

Unlike in the case of ordinary exchangeability, we need to consider two measurement procedures in which different decisions are made. Conditional probability models of decision problems express judgements that hold whatever decision is made. Under this interpretation, a conditional probability model $\mathbb{P}_{\square}^{Y|D}$ can tell us about the comparison of \mathcal{S} and \mathcal{S}' (see Section ??). In particular, it expresses the following attitude:

- If we decide on α , then the consequences are described by $\mathbb{P}_{\alpha}^D \mathbb{P}_{\square}^{Y|D}$
- If we decide on α' such that $\mathbb{P}_{\alpha'}^D = \mathbb{P}_{\alpha}^D$ then the consequences are described by $\mathbb{P}_{\alpha}^D \mathbb{P}_{\square}^{Y|D}$

Thus if we hold that \mathcal{S} is equivalent to \mathcal{S}' for all $\alpha \in A$, we conclude that \mathbb{P}_{\square} commutes with exchange.

There is a related kind of symmetry that involves shuffling “experimental units”. This is not an operation that can be described purely mathematically, but there may be a more intuitively appealing case available that this kind of shuffle yields an equivalent experiment. For example, if an “experimental unit” is a patient who receives a treatment and whose recovery is then followed up on as a consequences, we might be willing to regard an experiment equivalent if the order of patients in the experiment is shuffled. We think it is sometimes reasonable to deduce exchange commutativity from this kind of symmetry, but as we will see care is needed when doing so.

6.6 Example: exchange commutativity in the context of treatment choices

Consider the following two scenarios:

1. Dr Alice is going to see two patients who are both complaining of lower back pain. Prior to seeing either, she decides deterministically on $(\mathcal{D}_1, \mathcal{D}_2)$ for treatment decisions \mathcal{D}_1 and \mathcal{D}_2
2. As before, but \mathcal{D}_1 is chosen after examining patient 1, and \mathcal{D}_2 after examining patient 2

Alice could model either situation with a conditional probability model $(\mathbb{P}_{\square}^{Y_1 Y_2 | D_1 D_2}, A)$. In either situation, might the model be exchange commutative?

For each scenario, we want to consider two measurement procedures: first, the “original” measurement procedure, and then the measurement procedure with a swapped choices and consequences of interest. The question we want to consider is whether both procedures should be considered equivalent.

We will make two assumptions about measurement procedure equivalence: first, a measurement procedure is equivalent to an identical procedure in which patients are interchanged. Second, a measurement procedure is equivalent to an identical procedure in which the order of treatment and measurement of outcomes is interchanged.

We will describe measurement procedures using pseudocode, because this offers the opportunity to be precise about operations like swaps. Note that descriptions of measurement procedures, in pseudocode or otherwise, are incomplete descriptions.

Suppose the first scenario corresponds to the following procedure \mathcal{S} .

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

Our assumption of patient interchangeability means that the following procedure is equivalent

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

Consider the swapped procedure

```
procedure  $\mathcal{S}$ 
   $(\mathcal{D}_1, \mathcal{D}_2) \leftarrow \text{choose\_treatments}$ 
   $\mathcal{Y}_1 \leftarrow \text{apply}(\mathcal{D}_1, \text{patient A})$ 
   $\mathcal{Y}_2 \leftarrow \text{apply}(\mathcal{D}_2, \text{patient B})$ 
  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 measurement 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)
   $\alpha \leftarrow \text{choose\_alpha}$ 
   $(\mathcal{D}_1, \mathcal{D}_2) \leftarrow \text{decisions}(\alpha)$ 
   $\mathcal{Y}_2 \leftarrow \text{apply}(\mathcal{D}_2, \text{patient A})$ 
   $\mathcal{Y}_1 \leftarrow \text{apply}(\mathcal{D}_1, \text{patient B})$ 
  return  $(\alpha, \mathcal{D}_1, \mathcal{D}_2, \mathcal{Y}_1, \mathcal{Y}_2)$ 
end procedure
```

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)
   $\alpha \leftarrow \text{choose\_alpha}$ 
   $(\mathcal{D}_1, \mathcal{D}_2) \leftarrow \text{decisions}(\alpha)$ 
   $(\mathcal{D}_1^{\text{swap}}, \mathcal{D}_2^{\text{swap}}) \leftarrow (\mathcal{D}_2, \mathcal{D}_1)$ 
   $\mathcal{Y}_1^{\text{swap}} \leftarrow \text{apply}(\mathcal{D}_1^{\text{swap}}, \text{patient A})$ 
   $\mathcal{Y}_2^{\text{swap}} \leftarrow \text{apply}(\mathcal{D}_2^{\text{swap}}, \text{patient B})$ 
   $(\mathcal{Y}_1, \mathcal{Y}_2) \leftarrow (\mathcal{Y}_2^{\text{swap}}, \mathcal{Y}_1^{\text{swap}})$ 
  return  $(\alpha, \mathcal{D}_1, \mathcal{D}_2, \mathcal{Y}_1, \mathcal{Y}_2)$ 
```

end procedure

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

procedure \mathcal{S}''

assert(patient A knowledge=patient B knowledge)

$\alpha \leftarrow \text{choose_alpha}$

$(\mathcal{D}_1, \mathcal{D}_2) \leftarrow \text{decisions}(\alpha)$

$\mathcal{Y}_2 \leftarrow \text{apply}(\mathcal{D}_2, \text{patient A})$

$\mathcal{Y}_1 \leftarrow \text{apply}(\mathcal{D}_1, \text{patient B})$

return $(\alpha, \mathcal{D}_1, \mathcal{D}_2, \mathcal{Y}_1, \mathcal{Y}_2)$

end procedure

Thus \mathcal{S}'' is exactly the same as \mathcal{S}' , which by assumption is equivalent to the original \mathcal{S} , and so the assumptions of interchangeable patients and reversible order of treatment application imply the model should commute with exchange. Thus, if we could extend this example to an infinite sequence of patients, there would exist a Markov kernel $\mathbb{P}_{\square}^{Y|DH} : D \times H \rightarrow Y$ 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 $\text{decisions}(\alpha)$ 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 applying 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)

$\alpha \leftarrow \text{choose_alpha}$

patient A knowledge \leftarrow inspect(patient A)

patient B knowledge \leftarrow inspect(patient B)

$(\mathcal{D}_1, \mathcal{D}_2) \leftarrow \text{vagueDecisions}(\alpha, \text{patient A knowledge}, \text{patient B knowledge})$

$\mathcal{Y}_1 \leftarrow \text{apply}(\mathcal{D}_1, \text{patient A})$

$\mathcal{Y}_2 \leftarrow \text{apply}(\mathcal{D}_2, \text{patient B})$

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)

$\alpha \leftarrow \text{choose_alpha}$

patient B knowledge \leftarrow inspect(patient B)

patient A knowledge \leftarrow inspect(patient A)

$(\mathcal{D}_1, \mathcal{D}_2) \leftarrow \text{vagueDecisions}(\alpha, \text{patient B knowledge}, \text{patient A knowledge})$

$\mathcal{Y}_2 \leftarrow \text{apply}(\mathcal{D}_2, \text{patient A})$

$\mathcal{Y}_1 \leftarrow \text{apply}(\mathcal{D}_1, \text{patient B})$

```

    return ( $\alpha$ ,  $\mathcal{D}_1$ ,  $\mathcal{D}_2$ ,  $\mathcal{Y}_1$ ,  $\mathcal{Y}_2$ )
end procedure
While paired permutation of decisions and outcomes yields
procedure  $\mathcal{T}''$ 
    assert(patient A knowledge=patient B knowledge)
     $\alpha \leftarrow \text{choose\_alpha}$ 
    patient A knowledge  $\leftarrow$  inspect(patient A)
    patient B knowledge  $\leftarrow$  inspect(patient B)
    ( $\mathcal{D}_1$ ,  $\mathcal{D}_2$ )  $\leftarrow$  vagueDecisions( $\alpha$ , patient A knowledge, patient B knowledge)
     $\mathcal{Y}_2 \leftarrow \text{apply}(\mathcal{D}_2, \text{patient A})$ 
     $\mathcal{Y}_1 \leftarrow \text{apply}(\mathcal{D}_1, \text{patient B})$ 
    return ( $\alpha$ ,  $\mathcal{D}_1$ ,  $\mathcal{D}_2$ ,  $\mathcal{Y}_1$ ,  $\mathcal{Y}_2$ )
end procedure

```

\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

6.7 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 following:

- 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 6.11 (Deterministically controllable). Given a probability gap model $(\mathbb{P}_\square, \{\mathbb{P}_\alpha^D\}_A, f)$ on (Ω, \mathcal{F}) and a variable $X : \Omega \rightarrow X$, if for any $x \in X$ there exists $\alpha \in A$ such that $\mathbb{P}_\alpha^X = \delta_x$ then X is deterministically controllable.

As an example of this, suppose $X : \Omega \rightarrow X$ is a source of random numbers, the set of decisions D is a set of functions $X \rightarrow T$ for treatments $T : \Omega \rightarrow T$ and $W : \Omega \rightarrow W$ are the ultimate patient outcomes, with $Y_i = (W_i, T_i)$. Then it may be reasonable to assume that $W_i \perp\!\!\!\perp (D_i, X_i) | T_i H$ (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 $\mathbb{P}_{\square}^{Y_i | T_i H}$.

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.

6.8 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 $(\mathbb{P}_{\square}^{BY | D[M]}, \{\mathbb{P}_\alpha^D\}_A, f)$ 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 $\mathbb{P}_{\square}^{Y_i | B_i}$ exist? We have by Lemma 6.12 that this exists if and only if $Y_i \perp\!\!\!\perp_{\mathbb{P}_{\square}} D_i | B_i$. 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.

6.9 Inferred causal contractibility

Perhaps the simplest case can be made for *derived causal contractibility*. If we have a prior judgement of causal contractibility with respect to the pair of variables (D, Y) , and we can collect data in a regime in which D has full support, then we can derive additional causal contractibility properties from conditional independences observed in the data.

Theorem 6.12. *Given \mathbb{P}_C with decisions D and consequences (Y, X) , if \mathbb{P}_C is $(D; X, Y)$ -causally contractible with response conditional $\mathbb{P}_C^{X_1 Y_1 | D_1 H}$ and there exists $\alpha \in C$ with $\mathbb{P}_\alpha^D \gg \{\mathbb{P}_\beta^D | \beta \in C\}$ and $Y_i \perp\!\!\!\perp_{\mathbb{P}_\alpha} D_i | H X_i$ for all $i \in M$, then \mathbb{P}_C is also $(Y; X)$ -causally contractible.*

Proof. Sufficiency: We want to show that $Y_i \perp\!\!\!\perp_{\mathbb{P}_\alpha} (R V Y_{\{i\}^C}, X_{\{i\}^C}, D) | (H, X_i)$ for all $i \in M$. Then the result follows by noting that $\mathbb{P}_C^{Y | X H D}$ exists by taking a higher order conditional with respect to $\mathbb{P}_C^{X Y H | D}$ and $\mathbb{P}_C^{Y_i | X_i H}$ therefore exists by application of Corollary 5.28.

From causal contractibility we have

$$(X_i, Y_i) \perp\!\!\!\perp_{\mathbb{P}_\alpha} (X_{\{i\}^C}, Y_{\{i\}^C}, D_{\{i\}^C}) | H D_i \quad (180)$$

$$Y_i \perp\!\!\!\perp_{\mathbb{P}_\alpha} (Y_{\{i\}^C}, X_{\{i\}^C}) | H D_i X_i \quad (181)$$

Where Eq. 181 follows from 180 by weak union.

Thus by contraction, $Y_i \perp\!\!\!\perp_{\mathbb{P}_\alpha} (Y_{\{i\}^C}, X_{\{i\}^C}, D) | H X_i$.

By Corollary 5.28 and the existence of $\mathbb{P}^{Y_i X_i | H D_i}$ for all $i \in M$, $\mathbb{P}_\square^{Y_i | H X_i}$ exists for all i . Furthermore, because $\mathbb{P}^{Y_i X_i | H D_i} = \mathbb{P}^{Y_j X_j | H D_j}$ for all $i, j \in M$, $\mathbb{P}_\square^{Y_i | H X_i} = \mathbb{P}_\square^{Y_j | H X_j}$ 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

$$\mathbb{P}_\alpha^{Y_0 | X_0 D_0 H}(B | x, d, h) = \mathbb{P}_\alpha^{Y_0 | X_0 H}(B | x, h) \quad (182)$$

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

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

$$\mathbb{P}_{\alpha_d}^{Y_i | X_i H}(B | x, h) = \int_D \mathbb{P}_{\alpha_d}^{Y_i | X_i D_i H}(B | x, d', h) \delta_d(d d') \quad (183)$$

$$= \mathbb{P}_{\alpha_d}^{Y_i | X_i D_i H}(B | x, d, h) \quad (184)$$

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}_{α_d} measure 0.

Need to add independence of hypothesis to representation theorem

However, note that for any α

$$\mathbb{P}_\alpha^{\mathbf{X}_i \mathbf{H} \mathbf{D}_i}(E \times F \times G) = \sum_{d \in G} \mathbb{P}_\alpha^{\mathbf{D}_i}(d) \mathbb{P}_\square^{\mathbf{X}_i \mathbf{H} | \mathbf{D}_i}(E \times F | d) \quad (185)$$

$$= \sum_{d \in G} \mathbb{P}_\alpha^{\mathbf{D}_i}(d) \sum_{d' \in D} \mathbb{P}_\square^{\mathbf{X}_i \mathbf{H} | \mathbf{D}_i}(E \times F | d') \mathbb{P}_{\alpha_d}^{\mathbf{D}_i}(\{d'\}) \quad (186)$$

$$= \sum_{d \in G} \mathbb{P}_\alpha^{\mathbf{D}_i}(d) \mathbb{P}_{\alpha_d}^{\mathbf{X}_i \mathbf{H} \mathbf{D}_i}(E \times F \times \{d\}) \quad (187)$$

Thus for each $d \in D$ the set $\{d\} \times C \subset D \times X \times H$ is of uniform \mathbb{P}_α 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

$$\mathbb{P}_\square^{\mathbf{Y}_0 | \mathbf{X}_0 \mathbf{H}}(B | x, h) = \mathbb{P}_\square^{\mathbf{Y}_0 | \mathbf{X}_0 \mathbf{D}_0 \mathbf{H}}(B | x, d, h) \quad (188)$$

as desired. \square

7 Conclusion

Given a set of choices and the ability to compare the desirability of different outcomes, if we want 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 think 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 applicable 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 contractible with respect to (D_i, Y_i) pairs, if there exists some intermediate X_i such that $Y_i \perp\!\!\!\perp D_i | X_i$ then causal contractibility also holds with respect to (X_i, Y_i) pairs. This implies, for example, in a randomised experiment where the choices D_i are functions from a random source R_i to treatments X_i , we not only have response conditionals $\mathbb{P}_{\square}^{Y_i | D_i}$ that tell us how outcomes respond to treatment assignment functions, but also response conditionals $\mathbb{P}_{\square}^{Y_i | X_i}$ 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 for 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 $Y \perp\!\!\!\perp E | X$, we also have causal contractibility with respect to (X, Y) .

7.1 Choices aren’t always known

One area of potential 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 trialing 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 grant maker 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 to the extent that the decision theoretic approach aims to be applicable to many common causal inference questions, it must come with some understanding of how to deal with poorly specified choices. One feature of the probability set approach we can exploit is: if the set C of choices for our model \mathbb{P}_C contains the true set C^* of choices, then universal features of \mathbb{P}_C will also be universal features of \mathbb{P}_{C^*} as the latter is a subset of the former. Thus if there is uncertainty about the actual set of choices that we should be considering, we may still be able to posit a large set of choices that we believe will contain the true set of interest.

8 Appendix, needs to be organised

8.1 Existence of conditional probabilities

Lemma 8.1 (Conditional pushforward). *Suppose we have a sample space (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$, $Z : \Omega \rightarrow Z$ and a probability set $\mathbb{P}_{\{\}}^X$ with conditional $\mathbb{P}_{\{\}}^{X|Y}$ such that $Z = f \circ Y$ for some $f : Y \rightarrow Z$. Then there exists a conditional probability $\mathbb{P}_{\{\}}^{Z|X} = \mathbb{P}_{\{\}}^{Y|X} \mathbb{F}_f$.*

Proof. Note that $(X, Z) = (\text{id}_X \otimes f) \circ (X, Y)$. Thus, by Lemma 5.11, for any $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$

$$\mathbb{P}_\alpha^{\mathbf{XZ}} = \mathbb{P}_\alpha^{\mathbf{XY}} \mathbb{F}_{\text{id}_X \otimes f} \quad (189)$$

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

$$\mathbb{F}_{\text{id}_X \otimes f}(A \times B|x, y) = \delta_x(A) \delta_{f(y)}(B) \quad (190)$$

$$= \mathbb{F}_{\text{id}_X}(A|x) \otimes \mathbb{F}_f(B|y) \quad (191)$$

$$\implies \mathbb{F}_{\text{id}_X \otimes f} = \mathbb{F}_{\text{id}_X} \otimes \mathbb{F}_f \quad (192)$$

Thus

$$\mathbb{P}_\alpha^{\mathbf{XZ}} = (\mathbb{P}_\alpha^{\mathbf{X}} \odot \mathbb{P}_{\{\}}^{\mathbf{Y|X}}) \mathbb{F}_{\text{id}_X} \otimes \mathbb{F}_f \quad (193)$$

$$= \begin{array}{c} \text{X} \\ \curvearrowright \\ \triangleleft \mathbb{P}_\alpha^{\mathbf{X}} \text{---} \bullet \text{---} \square \mathbb{P}_{\{\}}^{\mathbf{Y|X}} \text{---} \square \mathbb{F}_f \text{---} \text{Z} \end{array} \quad (194)$$

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

Theorem 8.2 (Existence of regular conditionals). *Suppose we have a sample space (Ω, \mathcal{F}) , variables $\mathbf{X} : \Omega \rightarrow X$ and $\mathbf{Y} : \Omega \rightarrow Y$ with Y standard measurable and a probability model \mathbb{P}_α on (Ω, \mathcal{F}) . Then there exists a conditional $\mathbb{P}_\alpha^{\mathbf{Y|X}}$.*

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

Theorem 8.3 (Existence of higher order valid conditionals with respect to probability sets). *Suppose we have a sample space (Ω, \mathcal{F}) , variables $\mathbf{X} : \Omega \rightarrow X$ and $\mathbf{Y} : \Omega \rightarrow Y$, $\mathbf{Z} : \Omega \rightarrow Z$ and a probability set $\mathbb{P}_{\{\}}$ with regular conditional $\mathbb{P}_{\{\}}^{\mathbf{YZ|X}}$ and Y and Z standard measurable. Then there exists a regular $\mathbb{P}_{\{\}}^{\mathbf{Z|(Y|X)}}$.*

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

$$n(f^{-1}(y)|x, y) = 1 \quad (195)$$

$$m(\mathbf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x, y) m \mathbb{F}_f(dy|x) \forall A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z} \quad (196)$$

In particular, $\mathbb{P}_{\{\}}^{\mathbf{YZ|X}}$ is a Borel measurable map $X \rightarrow Y \times Z$. Thus equation 196 implies for all $A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z}$

$$\mathbb{P}_{\{\}}^{\mathbf{YZ}|\mathbf{X}}(\mathbf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x, y) \mathbb{P}_{\{\}}^{\mathbf{YZ}|\mathbf{X}} \mathbb{F}_f(dy|x) \quad (197)$$

$$= \int_A n(B|x, y) \mathbb{P}_{\{\}}^{\mathbf{Y}|\mathbf{X}}(dy|x) \quad (198)$$

Where Equation 198 follows from Lemma 8.1.

Then, for any $\mathbb{P}_\alpha \in \mathbb{P}_{\{\}}$

$$\mathbb{P}_{\{\}}^{\mathbf{YZ}|\mathbf{X}}(\mathbf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x, y) \mathbb{P}_\alpha^{\mathbf{Y}|\mathbf{X}}(dy|x) \quad (199)$$

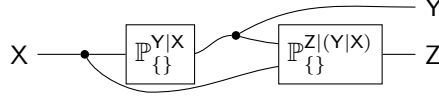
which implies n is a version of $\mathbb{P}_{\{\}}^{\mathbf{YZ}|\mathbf{X}}$. By Lemma 8.1, $n\mathbb{F}_f$ is a version of $\mathbb{P}_{\{\}}^{\mathbf{Z}|\mathbf{Y}|\mathbf{X}}$. \square

We might be motivated to ask whether the higher order conditionals in Theorem 8.3 can be chosen to be valid. Despite Lemma 8.8 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 195 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 (Ω, \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 8.4 (Higher order conditionals). *Suppose we have a sample space (Ω, \mathcal{F}) , variables $\mathbf{X} : \Omega \rightarrow X$ and $\mathbf{Y} : \Omega \rightarrow Y$, $\mathbf{Z} : \Omega \rightarrow Z$ and a probability set $\mathbb{P}_{\{\}}$ with conditional $\mathbb{P}_{\{\}}^{\mathbf{YZ}|\mathbf{X}}$. Then $\mathbb{P}_{\{\}}^{\mathbf{Z}|\mathbf{Y}|\mathbf{X}}$ is a version of $\mathbb{P}_{\{\}}^{\mathbf{Z}|\mathbf{Y}\mathbf{X}}$*

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



$$\mathbb{P}_\alpha^{YZ|X} = \quad (200)$$

$$\Rightarrow \mathbb{P}_\alpha^{XYZ} = \triangleleft \mathbb{P}_\alpha^X \quad (201)$$

$$= \quad (202)$$

$$= \quad (203)$$

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

Theorem 8.5. *Given probability gap model $\mathbb{P}_\{\}$, X, Y, Z such that $\mathbb{P}_\{\}^{Z|YX}$ exists, $\mathbb{P}_\{\}^{Z|Y}$ exists iff $Z \perp\!\!\!\perp_{\mathbb{P}_\{\}} X|Y$.*

Proof. If: If $Z \perp\!\!\!\perp_{\mathbb{P}_\{\}} X|Y$ then by Theorem 5.26, for each $\mathbb{P}_\alpha \in \mathbb{P}_\{\}$ there exists $\mathbb{P}_\alpha^{Z|Y}$ such that

$$\mathbb{P}_\alpha^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (204)$$

\square

Theorem 8.6 (Valid higher order conditionals). *Suppose we have a sample space (Ω, \mathcal{F}) , variables $X: \Omega \rightarrow X$ and $Y: \Omega \rightarrow Y$, $Z: \Omega \rightarrow Z$ and a probability set $\mathbb{P}_\{\}$ with regular conditional $\mathbb{P}_\{\}^{YZ|X}$, Y discrete and Z standard measurable. Then there exists a valid regular $\mathbb{P}_\{\}^{Z|XY}$.*

Proof. By Theorem 8.3, we have a higher order conditional $\mathbb{P}_\{\}^{Z|(Y|X)}$ which, by Theorem 8.4 is also a version of $\mathbb{P}_\{\}^{Z|XY}$.

Defining $\mathbf{O} := \text{id}_\Omega$ (the identity function $\Omega \rightarrow \Omega$), $\mu^{|\mathbf{X}}$ is a version of $\mu^{\mathbf{O}|\mathbf{X}}$. Note also that $\mathbf{Y} = \mathbf{Y} \circ \mathbf{O}$. Thus by Lemma 8.1, \mathbb{K} is a version of $\mu^{\mathbf{Y}|\mathbf{X}}$.

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

$$\mathbb{K}(\mathbf{Y} \bowtie A | x) = \delta_x(\mathbf{Y}^{-1}(A)) \quad (210)$$

$$= 0 \quad (211)$$

Thus \mathbb{K} is valid. \square

Theorem 8.9 (Validity). *Given (Ω, \mathcal{F}) , $\mathbf{X} : \Omega \rightarrow X$, $\mathbb{J} \in \Delta(X)$ with Ω and X standard measurable, there exists some $\mu \in \Delta(\Omega)$ such that $\mu^{\mathbf{X}} = \mathbb{J}$ if and only if \mathbb{J} is a valid distribution.*

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

$$\mu^{\mathbf{X}}(x) = \mu(\mathbf{X} \bowtie x) \quad (212)$$

$$= \sum_{x' \in X} \mathbb{J}(x') \mathbb{K}(\mathbf{X} \bowtie x | x') \quad (213)$$

$$= 0 \quad (214)$$

$$\neq \mathbb{J}(x) \quad (215)$$

\square

Lemma 8.10 (Semidirect product defines an intersection of probability sets). *Given (Ω, \mathcal{F}) , $\mathbf{X} : \Omega \rightarrow (X, \mathcal{X})$, $\mathbf{Y} : \Omega \rightarrow (Y, \mathcal{Y})$, $\mathbf{Z} : \Omega \rightarrow (Z, \mathcal{Z})$ all standard measurable and maximal probability sets $\mathbb{P}_{\{\}}^{\mathbf{Y}|\mathbf{X}[M]}$ and $\mathbb{Q}_{\{\}}^{\mathbf{Z}|\mathbf{YX}[M]}$ then defining*

$$\mathbb{R}_{\{\}}^{\mathbf{YZ}|\mathbf{X}} := \mathbb{P}_{\{\}}^{\mathbf{Y}|\mathbf{X}} \odot \mathbb{Q}_{\{\}}^{\mathbf{Z}|\mathbf{YX}} \quad (216)$$

we have

$$\mathbb{R}_{\{\}}^{\mathbf{YZ}|\mathbf{X}[M]} = \mathbb{P}_{\{\}}^{\mathbf{Y}|\mathbf{X}[M]} \cap \mathbb{Q}_{\{\}}^{\mathbf{Z}|\mathbf{YX}[M]} \quad (217)$$

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

$$\mathbb{R}_a^{\mathbf{XYZ}} = \mathbb{R}_a^{\mathbf{X}} \odot \mathbb{P}_{\{\}}^{\mathbf{Y}|\mathbf{X}} \odot \mathbb{Q}_{\{\}}^{\mathbf{Z}|\mathbf{YX}} \quad (218)$$

$$\implies \mathbb{R}_a^{\mathbf{XY}} = \mathbb{R}_a^{\mathbf{X}} \odot \mathbb{P}_{\{\}}^{\mathbf{Y}|\mathbf{X}} \quad (219)$$

$$\wedge \mathbb{R}_a^{\mathbf{XYZ}} = \mathbb{R}_a^{\mathbf{XY}} \odot \mathbb{Q}_{\{\}}^{\mathbf{Z}|\mathbf{YX}} \quad (220)$$

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

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

Lemma 8.11 (Equivalence of validity definitions). *Given $\mathbb{X} : \Omega \rightarrow X$, with Ω and X standard measurable, a probability measure $\mathbb{P}^{\mathbb{X}} \in \Delta(X)$ is valid if and only if the conditional $\mathbb{P}^{\mathbb{X}|\ast} := \ast \mapsto \mathbb{P}^{\mathbb{X}}$ is valid.*

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

$$\forall A \in \mathcal{X} : \mathbb{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathbb{X}|\ast}(A|\ast) = 0 \quad (221)$$

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

$$\forall A \in \mathcal{X} : \mathbb{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathbb{X}}(A) = 0 \quad (222)$$

\square

Lemma 8.12 (Semidirect product of valid candidate conditionals is valid). *Given (Ω, \mathcal{F}) , $\mathbb{X} : \Omega \rightarrow X$, $\mathbb{Y} : \Omega \rightarrow Y$, $\mathbb{Z} : \Omega \rightarrow Z$ (all spaces standard measurable) and any valid candidate conditional $\mathbb{P}^{\mathbb{Y}|\mathbb{X}}$ and $\mathbb{Q}^{\mathbb{Z}|\mathbb{Y}\mathbb{X}}$, $\mathbb{P}^{\mathbb{Y}|\mathbb{X}} \odot \mathbb{Q}^{\mathbb{Z}|\mathbb{Y}\mathbb{X}}$ is also a valid candidate conditional.*

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

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

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

$$\mathbb{R}^{\mathbb{Y}\mathbb{Z}|\mathbb{X}}(B \times C|x) = \int_B \mathbb{Q}^{\mathbb{Z}|\mathbb{Y}\mathbb{X}}(C|y, x) \mathbb{P}^{\mathbb{Y}|\mathbb{X}}(dy|x) \quad (223)$$

$$\leq \mathbb{P}^{\mathbb{Y}|\mathbb{X}}(B|x) \quad (224)$$

$$= 0 \quad (225)$$

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

$$\mathbb{R}^{\mathbb{Y}\mathbb{Z}|\mathbb{X}}(B \times C|x) = \int_B \mathbb{Q}^{\mathbb{Z}|\mathbb{Y}\mathbb{X}}(C|y, x) \mathbb{P}^{\mathbb{Y}|\mathbb{X}}(dy|x) \quad (226)$$

$$= 0 \quad (227)$$

\square

Corollary 8.13 (Valid conditionals are validly extendable to valid distributions). *Given Ω , $U : \Omega \rightarrow U$, $W : \Omega \rightarrow W$ and a valid conditional $\mathbb{T}^{W|U}$, then for any valid conditional \mathbb{V}^U , $\mathbb{V}^U \odot \mathbb{T}^{W|U}$ is a valid probability.*

Proof. Applying Lemma 8.12 choosing $X = *$, $Y = U$, $Z = W$ and $\mathbb{P}^{Y|X} = \mathbb{V}^{U|*}$ and $\mathbb{Q}^{Z|YX} = \mathbb{T}^{W|U*}$ we have $\mathbb{R}^{WU|*} := \mathbb{V}^{U|*} \odot \mathbb{T}^{W|U*}$ is a valid conditional probability. Then $\mathbb{R}^{WU} \cong \mathbb{R}^{WU|*}$ is valid by Theorem 8.11. \square

Theorem 8.14 (Validity of conditional probabilities). *Suppose we have Ω , $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, with Ω , X , Y discrete. A conditional $\mathbb{T}^{Y|X}$ is valid if and only if for all valid candidate distributions \mathbb{V}^X , $\mathbb{V}^X \odot \mathbb{T}^{Y|X}$ is also a valid candidate distribution.*

Proof. If: this follows directly from Corollary 8.13.

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

$$(\mathbb{V}^X \odot \mathbb{T}^{Y|X})(x, y) = \mathbb{T}^{Y|X}(y|x) \mathbb{V}^X(x) \quad (228)$$

$$= \mathbb{T}^{Y|X}(y|x) \quad (229)$$

$$> 0 \quad (230)$$

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

8.3 Conditional independence

Theorem 5.27. *Given standard measurable (Ω, \mathcal{F}) , variables $W : \Omega \rightarrow W$, $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ and a probability set \mathbb{P}_C with uniform conditional probability $\mathbb{P}_C^{Y|WX}$ and $\alpha \in C$ such that $\mathbb{P}_\alpha^{WX} \gg \{\mathbb{P}_\beta^{WX} | \beta \in C\}$, $Y \perp_{\mathbb{P}_\alpha} X|W$ if and only if there is a version of $\mathbb{P}_C^{Y|WX}$ and $\mathbb{K} : W \rightarrow Y$ such that*

$$\mathbb{P}_C^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (231)$$

Proof. If: By assumption, for every $\beta \in A$ we can write

$$\mathbb{P}_\beta^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (232)$$

And so, by Theorem 5.26, $Y \perp_{\mathbb{P}_\beta} X|W$ for all $\beta \in A$, and in particular $Y \perp_{\mathbb{P}_\alpha} X|W$. Only if: By Theorem 5.26, there exists a version of $\mathbb{P}_\alpha^{Y|WX}$ such that

$$\mathbb{P}_\alpha^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{P}_\alpha^{Y|W}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (233)$$

Because \mathbb{P}_α^{WX} dominates $\{\mathbb{P}_\beta^{WX} | \beta \in C\}$ and the set of points on which $\mathbb{P}_\alpha^{Y|WX}$ differs from $\mathbb{P}_C^{Y|WX}$ is of \mathbb{P}_α measure 0, this set must also be of \mathbb{P}_β measure 0 for all $\beta \in C$. Therefore $\mathbb{P}_\alpha^{Y|WX}$ is a version of $\mathbb{P}_C^{Y|WX}$, and so

$$\begin{array}{c} \mathbb{P}_C^{Y|WX} = W \text{ --- } \boxed{\mathbb{P}_\alpha^{Y|W}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (234)$$

□

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 $Y \perp_{\mathbb{P}_\square} X$ because X is deterministic for any $\alpha \in A$. But $\mathbb{P}_\square^{Y|X}$ is not necessarily unresponsive to X .

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

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

Proof.

$$\forall x : f(x) \implies g(x) \quad \text{premise} \quad (235)$$

$$\forall x : f(x) \quad \text{premise} \quad (236)$$

$$f(a) \quad \text{UI on 236 sub } a/x \quad (237)$$

$$f(a) \implies g(a) \quad \text{UI on 235 sub } a/x \quad (238)$$

$$g(a) \quad \text{MP 237 and 238} \quad (239)$$

$$\forall x : g(x) \quad \text{UG on 239} \quad (240)$$

$$(\forall x : f(x)) \implies (\forall x : g(x)) \quad \text{CP 236 – 240} \quad (241)$$

$$[\forall x : (f(x) \implies g(x))] \implies [(\forall x : f(x)) \implies (\forall x : g(x))] \quad \text{CP 235–241} \quad (242)$$

Where UI: universal instantiation, UG: universal generalisation, MP: modus ponens and CP: conditoinal proof. With thanks to (StackExchange) for the proof. □

8.4 Maximal probability sets and valid conditionals

We have defined probability sets and uniform conditional probabilities. Thus, if we start with a probability set, we know how to check if certain uniform conditional probabilities exist or not. However, there is a particular line of reasoning that comes up most often in the graphical models tradition of causal inference where we start with collections of conditional probabilities and assemble them into probability models as needed. A simple example of this is the causal Bayesian network given by the graph $X \longrightarrow Y$ and some observational probability distribution $\mathbb{P}^{XY} \in \Delta(X \times Y)$. Using the standard notion of “hard interventions on X ”, this model induces a probability set which we could informally describe as the set $\mathbb{P}_{\square} := \{\mathbb{P}_a^{XY} | a \in X \cup \{*\}\}$ where $*$ is a special element corresponding to the observational setting. The graph $X \longrightarrow Y$ implies the existence of the uniform conditional probability $\mathbb{P}_{\square}^{Y|X}$ under the nominated set of interventions, while the usual rules of hard interventions imply that $\mathbb{P}_a^X = \delta_a$ for $a \in X$.

Reasoning “backwards” like this – from uniform conditionals and marginals back to probability sets – must be done with care. The probability set associated with a collection of conditionals and marginals may be empty or nonunique. Uniqueness may not always be required, but an empty probability set is clearly not a useful model.

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

We introduce the notion of *valid distributions* and *valid conditionals*. The key result here is: probability sets defined by collections of recursive valid conditionals and distributions are nonempty. While we suspect this condition is often satisfied by causal models in practice, we offer one example in the literature where it apparently is not. 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). As that work shows, there are many questions of this nature that can be asked and that are not addressed by the criterion of validity.

There is also a connection between the notion of validity and the notion of *unique solvability* in Bongers et al. (2016). We ask “when can a set of conditional probabilities together with equations be jointly satisfied by a probability model?” while Bongers et. al. ask when a set of equations can be jointly satisfied by a probability model.

Definition 8.15 (Valid distribution). Given (Ω, \mathcal{F}) and a variable $X : \Omega \rightarrow X$, an X -valid probability distribution is any probability measure $\mathbb{K} \in \Delta(X)$ such that $X^{-1}(A) = \emptyset \implies \mathbb{K}(A) = 0$ for all $A \in \mathcal{X}$.

Definition 8.16 (Valid conditional). Given (Ω, \mathcal{F}) , $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ a $Y|X$ -valid conditional probability is a Markov kernel $\mathbb{L} : X \rightarrow Y$ that assigns

probability 0 to impossible events, unless the argument itself corresponds to an impossible event:

$$\forall B \in \mathcal{Y}, x \in X : (X, Y) \bowtie \{x\} \times B = \emptyset \implies (\mathbb{L}(B|x) = 0) \vee (X \bowtie \{x\} = \emptyset) \quad (243)$$

Definition 8.17 (Maximal probability set). Given $(\Omega, \mathcal{F}), X : \Omega \rightarrow X, Y : \Omega \rightarrow Y$ and a $Y|X$ -valid conditional probability $\mathbb{L} : X \rightarrow Y$ the maximal probability set $\mathbb{P}_{\{\}}^{Y|X[M]}$ associated with \mathbb{L} is the probability set such that for all $\mathbb{P}_{\alpha} \in \mathbb{P}_{\{\}}, \mathbb{L}$ is a version of $\mathbb{P}_{\alpha}^{Y|X}$.

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

Lemma 8.12 shows that the semidirect product of any pair of valid conditional probabilities is itself a valid conditional. Suppose we have some collection of $X_i|X_{[i-1]}$ -valid conditionals $\{\mathbb{P}_i^{X_i|X_{[i-1]}} | i \in [n]\}$; then recursively taking the semidirect product $\mathbb{M} := \mathbb{P}_1^{X_1} \odot (\mathbb{P}_2^{X_2|X_1} \odot \dots)$ yields a $X_{[n]}$ 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 $\mathbb{K} : X \rightarrow Y$, \mathbb{K} might be an invalid $Y|X$ 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 $\mathbb{P}^{Y|X}$. On the other hand, if we want to take the product of \mathbb{K} with arbitrary valid X probabilities, then the validity of \mathbb{K} is necessary (Theorem 8.14).

Example 8.18. 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 $B = \frac{W}{H^2}$.

With this background, suppose we postulate a decision model in which body mass index can be directly controlled by a variable C , while height and weight are not. Specifically, we have a probability set \mathbb{P}_{\square} with

$$\mathbb{P}_{\square}^{B|WHC} = \begin{array}{c} H \text{ ---} * \\ C \text{ -----} B \\ W \text{ ---} * \end{array} \quad (244)$$

Then pick some $w, h, x \in \mathbb{R}$ such that $\frac{w}{h^2} \neq x$ and $(W, H) \bowtie (w, h) \neq \emptyset$ (which is to say, our measurement procedure could potentially yield (w, h) for a person’s

height and weight). We have $\mathbb{P}_{\square}^{\mathbf{B}|\mathbf{WHC}}(\{x\}|w, h, x) = 1$, but

$$(\mathbf{B}, \mathbf{W}, \mathbf{H}) \bowtie \{(x, w, h)\} = \{\omega | (\mathbf{W}, \mathbf{H})(\omega) = (w, h), \mathbf{B}(\omega) = \frac{w}{h^2}\} \quad (245)$$

$$= \emptyset \quad (246)$$

so $\mathbb{P}_{\square}^{\mathbf{B}|\mathbf{WHC}}$ is invalid. Thus there is some valid $\mu^{\mathbf{WHC}}$ such that the probability set $\mathbb{P}_{\square}^{\mathbf{B}|\mathbf{WHC}} = \mu^{\mathbf{WHC}} \odot \mathbb{P}_{\square}^{\mathbf{Y}|\mathbf{X}}$ is empty.

Validity rules out conditional probabilities like 244. We conjecture that in many cases this condition is implicitly taken into account – it is obviously silly to posit a model in which body mass index can be controlled independently of height and weight. We note, however, that presuming the authors intended their model to be interpreted according to the usual semantics of causal Bayesian networks, the invalid conditional probability 244 would be used to evaluate the causal effect of body mass index in the causal diagram found in Shahar (2009).

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 Validity. We thank Esther Duflo for her leadership on the handbook and 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 Evaluations 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.
- Stephan Bongers, Jonas Peters, Bernhard Schölkopf, and Joris M. Mooij. Theoretical Aspects of Cyclic Structural Causal Models. *arXiv:1611.06221 [cs, stat]*, November 2016. URL <http://arxiv.org/abs/1611.06221>. arXiv: 1611.06221.

- 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.
- 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. Beware of the DAG! In *Causality: Objectives and Assessment*, pages 59–86, February 2010. URL <http://proceedings.mlr.press/v6/dawid10a.html>.
- Philip Dawid. Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9(1):39–77, January 2021. ISSN 2193-3685. doi: 10.1515/jci-2020-0008. URL <https://www.degruyter.com/document/doi/10.1515/jci-2020-0008/html>. Publisher: De Gruyter.
- 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. Interditions, France, 1979. ISBN 978-2-7296-0030-3. INIS Reference Number: 13648743.

- Brendan Fong. Causal Theories: A Categorical Perspective on Bayesian Networks. *arXiv:1301.6201 [math]*, January 2013. URL <http://arxiv.org/abs/1301.6201>. arXiv: 1301.6201.
- Tobias Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, August 2020. ISSN 0001-8708. doi: 10.1016/j.aim.2020.107239. URL <https://www.sciencedirect.com/science/article/pii/S0001870820302656>.
- Sander Greenland and James M Robins. Identifiability, Exchangeability, and Epidemiological Confounding. *International Journal of Epidemiology*, 15(3): 413–419, September 1986. ISSN 0300-5771. doi: 10.1093/ije/15.3.413. URL <https://doi.org/10.1093/ije/15.3.413>.
- D. Heckerman and R. Shachter. Decision-Theoretic Foundations for Causal Reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, December 1995. ISSN 1076-9757. doi: 10.1613/jair.202. URL <https://www.jair.org/index.php/jair/article/view/10151>.
- M. A. Hernán and S. L. Taubman. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity*, 32(S3):S8–S14, August 2008. ISSN 1476-5497. doi: 10.1038/ijo.2008.82. URL <https://www.nature.com/articles/ijo200882>.
- 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.
- 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>.
- 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.
- P. Selinger. A Survey of Graphical Languages for Monoidal Categories. In Bob Coecke, editor, *New Structures for Physics*, Lecture Notes in Physics, pages 289–355. Springer, Berlin, Heidelberg, 2011. ISBN 978-3-642-12821-9. doi: 10.1007/978-3-642-12821-9_4. URL https://doi.org/10.1007/978-3-642-12821-9_4.
- 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.
- KyleW (StackExchange). 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).
- 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=tvtp&paperid=4710&option_lang=eng.
- Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.

Appendix: