

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

David Johnston, Cheng Soon Ong, Robert C. Williamson

March 22, 2022

Contents

1	Introduction	2
1.1	Our approach	4
1.2	Contributions	6
2	Variables in probabilistic models	7
2.1	Measurement procedures	9
2.2	Observable variables	10
2.3	Model variables	11
2.4	Variable sequences	12
2.5	Decision procedures	12
3	Decision problems	12
3.1	Other decision theoretic causal models	13
4	Probability prerequisites	14
4.1	The roles of variables and probabilistic models	14
4.2	Standard probability theory	15
4.3	Not quite standard probability theory	17
5	String diagram notation	18
5.1	Products	18
5.2	Elements of string diagrams	19
5.3	Iterated copy maps and plates	20
5.3.1	Examples	21
6	Probability sets	22
6.1	Semidirect product and almost sure equality	23
6.2	Conditional independence	25
6.3	Uniform conditional independence	27

7	When do response conditionals exist?	30
7.1	Sequential decision models	31
7.2	Causal contractibility	32
7.3	Existence of response conditionals	35
7.4	Example: backdoor adjustment	37
7.5	Assessing causal contractibility	37
7.6	Body mass index revisited	40
7.7	Weakening causal contractibility	40
8	Conclusion	41
8.1	Choices aren't always known	42
9	Appendix, needs to be organised	43
9.1	Markov categories	43
9.2	Existence of conditional probabilities	44
9.3	Validity	47
9.4	Conditional independence	51
9.5	Maximal probability sets and valid conditionals	51
9.6	Causal contractibility	54
9.7	Body mass index revisited	59

Abstract

Popular causal inference frameworks are missing key ingredients. Potential outcomes has no notion of manipulation, and so can only offer informal explanations of critical assumptions like “stable unit-treatment values” (SUTVA). Approaches that take manipulation as basic miss the fact that the basic role of a causal model is to represent uncertain knowledge of the consequences of different choices. As a result, they leave us searching in vain for “elementary manipulations” and prevent us from understanding how “causal effects” arise from symmetries in causal models. Incorporating both of these ingredients leads us to the “decision theoretic” approach to causal inference – causal models map sets of choices to probability distributions representing our knowledge of the consequences of these choices. With this perspective in hand, we offer a necessary and sufficient condition for the existence of “causal effects”, and show how the neglected condition of *permutability of identifiers* is a critical requirement for classical randomised experiments to support causal effects.

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 (CBNs) and Structural Causal Models (SCMs), provide a set of *intervention* operations that take probability distributions and a graph and return an *interventional 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 represent a potential or *counterfactual* value (Rubin, 2005).

It is generally accepted that unique interventional probability distributions or potential outcomes are inappropriate for some situations that which we might want causal models for. The potential outcomes framework discusses this in terms of the “stable unit-treatment value assumption” (SUTVA), which is offered as a necessary condition for the existence of potential outcomes, although we are not aware of any formal statements of SUTVA nor any proofs of necessity or sufficiency Rubin (2005); Imbens and Rubin (2015). A very similar issue has been considered in the graphical models community, and a number of different solutions have been proposed.

In the graphical models community, the problem of non-unique interventional distributions is sometimes framed as the problem of determining which variables are *causal variables*. While the distinction between causal variables and random variables is rarely made explicit, the two kinds of variable are not the same.

Two brief examples illustrate this point. First, *height in centimetres* H and *height in metres* H' are distinct random variables – concretely, $H = 100 * H'$. However it is impossible to take an action that fixes the value of one variable independently of the other. Thus if causal relationships require actions that correspond to the intervention operation in a directed acyclic graph, H and H' cannot have a causal relationship (this example is due to Eberhardt (2022)).

Second, a random variable X fails to be a causal variable with respect to Y – and fails to have corresponding potential outcomes Y^X – when there are multiple plausible actions that affect the value of X that are likely to have different consequences with respect to Y . Hernán and Taubman (2008) observed that many epidemiological papers have been published estimating the “causal effect” of body mass index B on mortality M . However, as Hernán and Taubman point out, body mass index may be altered by diet, exercise or surgery, and all three different choices are likely to have different consequences with respect to mortality M . That is, the consequences with regard to M are underdetermined by simply stipulating that an action has some effect on B . A very similar example is explored by Spirtes and Scheines (2004); Eberhardt (2022) who discuss the effect of cholesterol on heart disease, which they argue is similarly underdetermined.

In a response to Hernán and Taubman, Shahar (2009) argued that a properly specified intervention on body mass index will yield the conclusion that any “intervention on body mass index” must have no effect at all on mortality, because the causal effects are “fully attributable to confounding by weight and perhaps height”. This claim is supported by the use of a causal diagram in which weight W and height H are the causal parents of body mass index B . However, this prescription runs afoul of the problem our first example illustrated: one cannot alter body mass index independently of weight and height. For questions like this, Spirtes and Scheines (2004) suggest that we should say the effect of ambiguous manipulations cannot be resolved from the data, or to return multiple possible answers for this effect.

At least three responses to this problem can be found: Spirtes and Scheines

(2004); Eberhardt (2022); Chalupka et al. (2017) all resolve the ambiguity by appealing to a fundamental set of manipulations, and posit that causal variable pairs are those whose probabilistic relationships do not depend (in a particular sense) on which intervention is chosen. Woodward (2016) acknowledges the difficulty and reports that he was unable to provide a general characterisation of “well-defined interventions”. Finally, Hernán (2016) suggests that causal effects should be considered well-defined if sufficiently precise descriptions of an intervention are provided, as judged by consensus of experts.

1.1 Our approach

An observation previously made by Dawid (2021) and Heckerman and Shachter (1995) is that, when we are faced with a decision problem, the difficult question of finding “a fundamental set of well-defined interventions” is automatically resolved. In short: if we are facing a decision problem, then there must be a number of different choices that we want to compare. We need to compare every choice available, and there is no comparison that we need to make involving anything but the choices available. Thus, given a decision problem, the “fundamental set of interventions” is precisely the set of choices that require comparison.

We assume that the reason to construct causal models is to help solve decision problems, and that we therefore have a fundamental set of choices C to be compared. For each choice $c \in C$, we assume that some measurement procedure is carried out yielding results in a sample space (Ω, \mathcal{F}) , and we represent uncertain knowledge about the result of this procedure with a probability distribution on (Ω, \mathcal{F}) . The model postulated is a “probability function”; a map from C to probability distributions on a sample space Ω . Probability functions are also the kind of model used in many interventional causal models Pearl (2009); Richardson and Robins (2013), as well as the kind of model used in less widespread decision theoretic or Bayesian causal models Dawid (2021); Lattimore and Rohde (2019a).

“decision theoretic causal model” and an “interventional causal model” is a difference of interpretation. . However, they do lead to differences in the way that we construct models for a given problem.

For example, interventional causal models are typically models of what Dawid (2021) calls “generic variables”. If we have a sequence of independent and identically distributed variables $(X_1, Y_1), (X_2, Y_2), \dots$ that are associated with a concrete measurement procedure, sometimes the individual variables are replaced by a generic “ X ” and “ Y ”. The assumption that the sequences is independent and identically distributed usually doesn’t even make sense for a probability function modelling this sequence. Therefore, we are not in a rush to substitute variables as we typically understand them with generic variables.

While the difference is “merely” on the level of interpretation, it informs the way we use probability functions to construct causal models.

However, this is not enough on its own to resolve the question of when causal effects exist. The “causal effect of X on Y ” is not just some probabilistic function from the range of X to the range of Y . Rather, as typically understood, it has a few properties:

- It is a probabilistic function from the range of X to the range of Y
- It may be unknown prior to seeing any data but becomes known with certainty in the limit of infinite causally sufficient data
- It represents the “distribution of Y , given X ” no matter which choice is selected

The question we focus on here is, from the decision theoretic starting point, when can we talk in the usual manner about “the causal effect” of one variable X on another variable Y ? In order to answer a question like this, we need to be more specific about what a “causal effect” is. Our provisional definition of a causal effect is similar to an earlier analysis by Bruno De Finetti; De Finetti asked what we could possibly mean when we said a sequence of coin flips was distributed according to a “constant but unknown probability \mathbb{Q} ” de Finetti ([1937] 1992). Similarly, we take “a causal effect of X on Y ” to be

For an example, consider a collection of “interventional probabilities” $\mathbb{P}(Y|do(X = x))$ (Pearl, 2009, chap. 1.3). Such a collection may or may not be a model of a decision problem. As is typically understood, $\mathbb{P}(Y|do(X = x))$ is not known prior to seeing data, and may not even be known after seeing a large amount of data. Finally, $\mathbb{P}(Y|do(X = x))$ is often interpreted as “the probability distribution of Y , if I were to take some action that sets X to be equal to x ”. Usually implicit in this interpretation is that the model assigns Y the same probability distribution *whatever* action is taken that sets X equal to x . This principle is not always implicit – Chalupka et al. (2017) makes this an explicit requirement for a variable X to qualify as a “causal” variable with respect to Y .

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”, and *decision procedures* which are collections of measurement procedures indexed by the different possible choices we have available. Executing a measurement procedure involves interacting with the real world such that a unique element of a well-defined mathematical set is returned. Each measurement procedure in a decision procedure yields an element of the same set.

Because measurement procedures have mathematical sets as their “codomain”, functions can be composed with measurement procedures. Because their “domain” is the real world, we cannot perform composition in the reverse direction – measurement procedures cannot be composed with functions. We would prefer to work with functions than with measurement procedures, so we invoke a single complete measurement procedure that includes all of the different measurements we’re interested in for a particular problem. Individual measurements are obtained by composing functions with the complete measurement procedure. In this manner, each individual measurement is associated with a mathematical function, and these mathematical functions are our *variables*.

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 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 can contain 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 measurement procedures is somewhat nonstandard, 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 measurement 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 measurement procedure, while the second is a definition of a variable. Variables model procedures, but they are not the same thing. It is common, for instance, for statistical models to contain a mixture of observed and unobserved variables. Unobserved variables, by definition, are not associated with any measurement procedure.

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 .

2.1 Measurement procedures

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

We adopt the convention that the procedure name \mathcal{B} and the set of values B share the same letter.

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

Definition 2.3 (Equivalence of procedures). Two procedures \mathcal{B} and \mathcal{C} are equal if they both take values in X and $\mathcal{B} \bowtie x \iff \mathcal{C} \bowtie x$ for all $x \in X$.

If two involve different measurement actions in the real world but necessarily yield the same result, we say they are equivalent.

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

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

Definition 2.4 (Composition of functions with procedures). Given a procedure \mathcal{B} that takes values in some set B , and a function $f : B \rightarrow C$, 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:

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

$$\iff \mathcal{B} \text{ yields } f^{-1}(g^{-1}(x)) \quad (2)$$

$$\iff f \circ \mathcal{B} \text{ yields } g^{-1}(x) \quad (3)$$

$$\iff g \circ (f \circ \mathcal{B}) \text{ yields } x \quad (4)$$

One might wonder whether there is also some kind of “tensor product” 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 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} .

One could also consider measurement procedures to produce results in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (i.e. the reals with the Borel sigma-algebra) or a set isomorphic to it. This choice is often made in practice, and following standard practice we also often consider variables to take values in sets isomorphic to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. However, for measurement in particular this seems to be a choice of convenience rather than necessity – for any measurement with finite precision and range, it is possible to specify a finite set of possible results.

2.2 Observable variables

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

Definition 2.5 (Observable variable). Given a measurement procedure \mathcal{S} taking values in (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 2.6 (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 ”.

2.3 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 2.7 (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 $s \in \Psi$.

2.4 Variable sequences

Given $Y : \Omega \rightarrow X$, 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.

2.5 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, and trying to include the process for making a decision in the model creates some difficult problems.

We avoid these problems by not including the procedure for making a decision. 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.

Definition 2.8 (Decision procedure). A decision procedure is a collection $\{S_\alpha\}_{\alpha \in A}$ of measurement procedures. By convention, we label sub-procedures with the same subscript $\mathcal{X}_\alpha = X \circ S_\alpha$.

Like measurement procedures, a decision procedure $\{S_\alpha\}_{\alpha \in A}$ isn't a well-defined mathematical object; it's a "set" containing real-world actions. However, we think the meaning is clear enough to work with.

3 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 \mathcal{S}_α associated with the choice $\alpha \in C$ (see Section 2.5). We suppose that each \mathcal{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 \mathcal{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 procedure is then a set of probability distributions $\mathbb{P}_C := \{\mathbb{P}_\alpha | \alpha \in C\}$, which we call a *probability set*.

3.1 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 6.15, on the basis of Theorem 6.17.

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 6.3), 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*.

4 Probability prerequisites

Notation table, including iverson bracket

4.1 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 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 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.

4.2 Standard probability theory

Definition 4.1 (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 4.2 (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 4.3 (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 4.4 (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 4.5 (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 4.6 (Variable). Given a measureable 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 4.7 (Sequence of variables). Given a measureable 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 4.8 (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 4.9 (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

$$\mu\mathbb{K}(A) := \int_X \mathbb{K}(A|x)\mu(dx) \quad (5)$$

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

Definition 4.10 (Kernel-kernel products). Given (X, \mathcal{X}) , (Y, \mathcal{Y}) , (Z, \mathcal{Z}) and Markov kernels $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : Y \rightarrow Z$, $\mathbb{K}\mathbb{L}$ is a Markov kernel $X \rightarrow Z$ defined by

$$\mathbb{K}\mathbb{L}(A|x) := \int_Y \mathbb{L}(A|y)\mathbb{K}(dy|x) \quad (6)$$

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

Lemma 4.11 (Marginal distribution as a kernel product). *Given a probability space $(\mu, \Omega, \mathcal{F})$ and a variable $X : \Omega \rightarrow (X, \mathcal{X})$, define $\mathbb{F}_X : \Omega \rightarrow X$ by $\mathbb{F}_X(A|\omega) = \delta_{X(\omega)}(A)$, then*

$$\mu^X = \mu\mathbb{F}_X \quad (7)$$

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

$$\mu\mathbb{F}_X(A) = \int_\Omega \delta_{X(\omega)}(A)d\mu(\omega) \quad (8)$$

$$= \int_{X^{-1}(A)} d\mu(\omega) \quad (9)$$

$$= \mu^X(A) \quad (10)$$

□

4.3 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 4.12 (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 4.13 (Conditional distribution). Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, the probability of Y given X is any Markov kernel $\mu^{Y|X} : X \rightarrow Y$ such that

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

$$\iff \quad (12)$$

$$\mu^{XY} = \begin{array}{c} \text{X} \\ \curvearrowright \\ \triangleleft \mu^X \end{array} \begin{array}{c} \bullet \\ \text{---} \mu^{Y|X} \end{array} \text{Y} \quad (13)$$

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

Definition 4.14 (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 $\mu^{Y|X}$,

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

$$\iff \quad (15)$$

$$\mu^{ZY|X} = \begin{array}{c} \text{Y} \\ \curvearrowright \\ \text{X} \text{---} \bullet \end{array} \begin{array}{c} \text{---} \mu^{Y|X} \end{array} \begin{array}{c} \bullet \\ \text{---} \mu^{Z|(Y|X)} \end{array} \text{Z} \quad (16)$$

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 holds for conditional with respect to probability sets, which we will introduce later (Theorem 9.5).

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 9.4 for the extension of this theorem to probability sets.

5 String diagram notation

We make use of a string diagram notation for probabilistic reasoning. Graphical models are often employed in causal reasoning, and string diagrams are a kind of graphical notation for representing Markov kernels. The notation comes from the study of Markov categories, which are abstract categories that represent models of the flow of information. For our purposes, we don't use abstract Markov categories but instead focus on the concrete category of Markov kernels on standard measurable sets.

A coherence theorem exists for string diagrams and Markov categories. Applying certain transformations such as planar deformation or any of the commutative comonoid axioms to a string diagram yields an equivalent string diagram. The coherence theorem establishes that any proof constructed using string diagrams in this manner corresponds to a proof in any Markov category (Selinger, 2011). More comprehensive introductions to Markov categories can be found in Fritz (2020); Cho and Jacobs (2019).

5.1 Products

On discrete sets, probability measures are vectors and Markov kernels are matrices. Thus given a probability measure $\mu \in \Delta(X)$ and a Markov kernel $\mathbb{K} : X \rightarrow Y$, the product $\mu\mathbb{K} \in \Delta(Y)$ is a standard vector-matrix product. This idea generalises to measures and Markov kernels in general.

Definition 5.1 (measure-kernel product). Given a probability measure $\mu \in \Delta(X)$ and a Markov kernel $\mathbb{K} : X \rightarrow Y$, the product $\mu\mathbb{K} \in \Delta(Y)$ is a probability measure such that, for all $A \in \mathcal{Y}$

$$\mu\mathbb{K}(A) = \int_X \mathbb{K}(A|x)\mu(dx) \quad (17)$$

Definition 5.2 (kernel-kernel product). Given Markov kernels $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : Y \rightarrow Z$, the product $\mathbb{K}\mathbb{L} : X \rightarrow Z$ is a Markov kernel such that, for all $x \in X$ and $B \in \mathcal{Z}$

$$\mathbb{K}\mathbb{L}(B|x) = \int_Y \mathbb{L}(B|y)\mathbb{K}(dy|x) \quad (18)$$

We can also define a tensor product of kernels.

Definition 5.3 (Tensor product of kernels). Given $\mathbb{K} : X \rightarrow Y$ and $\mathbb{M} : W \rightarrow Z$, $\mathbb{K} \otimes \mathbb{M} : X \times W \rightarrow Y \times Z$ is given by

$$\mathbb{K} \otimes \mathbb{M}(A \times B|x, w) = \mathbb{K}(A|x)\mathbb{M}(B|w) \quad (19)$$

for all $A \in \mathcal{X}$, $B \in \mathcal{Y}$, $(x, w) \in X \times W$, and this uniquely defines $\mathbb{K} \otimes \mathbb{M}$.

5.2 Elements of string diagrams

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

$$\mathbb{K} := \text{---} \boxed{\mathbb{K}} \text{---} \quad (20)$$

$$\mu := \triangleleft \boxed{\mathbb{P}} \text{---} \quad (21)$$

Given two Markov kernels $\mathbb{L} : X \rightarrow Y$ and $\mathbb{M} : Y \rightarrow Z$, the product $\mathbb{L}\mathbb{M}$ is represented by drawing them side by side and joining their wires:

$$\mathbb{L}\mathbb{M} := X \text{---} \boxed{\mathbb{K}} \text{---} \boxed{\mathbb{M}} \text{---} Z \quad (22)$$

Given kernels $\mathbb{K} : W \rightarrow Y$ and $\mathbb{L} : X \rightarrow Z$, the tensor product $\mathbb{K} \otimes \mathbb{L} : W \times X \rightarrow Y \times Z$ is graphically represented by drawing kernels in parallel:

$$\mathbb{K} \otimes \mathbb{L} := \begin{array}{c} W \text{---} \boxed{\mathbb{K}} \text{---} Y \\ X \text{---} \boxed{\mathbb{L}} \text{---} Z \end{array} \quad (23)$$

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 $(\text{id}_X)(A|x) = \delta_x(A)$ for all $x \in X$, $A \in \mathcal{X}$, is a bare line:

$$\text{id}_X := X \text{---} X \quad (24)$$

Given some 1-element set $\{*\}$, the erase map $\text{del}_X : X \rightarrow \{*\}$ defined by $(\text{del}_X)(*|x) = 1$ for all $x \in X$ is a Markov kernel that “discards the input”. It looks like a lit fuse:

$$\text{del}_X := \text{---} * \text{---} X \quad (25)$$

The copy map $\text{copy}_X : X \rightarrow X \times X$ defined by $(\text{copy}_X)(A \times B|x) = \delta_x(A)\delta_x(B)$ for all $x \in X$, $A, B \in \mathcal{X}$ is a Markov kernel that makes two identical copies of the input. It is drawn as a fork:

$$\text{copy}_X := X \text{---} \text{---} \begin{array}{c} X \\ X \end{array} \quad (26)$$

The swap map $\text{swap}_{X,Y} : X \times Y \rightarrow Y \times X$, defined by $(\text{swap}_{X,Y})(A \times B | x, y) = \delta_x(B) \delta_y(A)$ for $(x, y) \in X \times Y$, $A \in \mathcal{X}$ and $B \in \mathcal{Y}$, swaps two inputs and is represented by crossing wires:

$$\text{swap}_{X,Y} := \begin{array}{c} \diagup \quad \diagdown \\ \diagdown \quad \diagup \end{array} \quad (27)$$

Diagrams in Markov categories satisfy the commutative comonoid axioms (see Definition 9.1)

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \quad \searrow \\ \nearrow \quad \searrow \end{array} = \text{---} \bullet \begin{array}{l} \nearrow \quad \searrow \\ \searrow \quad \nearrow \end{array} \end{array} \quad (28)$$

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \quad \searrow \\ \searrow \quad \nearrow \end{array} \quad \text{---} \quad \text{---} \bullet \begin{array}{l} \nearrow \quad \searrow \\ \searrow \quad \nearrow \end{array} \end{array} = \begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \quad \searrow \\ \searrow \quad \nearrow \end{array} \quad \text{---} \bullet \begin{array}{l} \nearrow \quad \searrow \\ \searrow \quad \nearrow \end{array} \end{array} \quad (29)$$

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \quad \searrow \\ \searrow \quad \nearrow \end{array} = \text{---} \bullet \begin{array}{l} \nearrow \quad \searrow \\ \searrow \quad \nearrow \end{array} \end{array} \quad (30)$$

as well as compatibility with the monoidal structure

$$\begin{array}{c} X \otimes Y \text{---} \bullet \quad X \text{---} \bullet \\ \text{---} \bullet \quad X \text{---} \bullet \end{array} = \begin{array}{c} X \text{---} \bullet \\ \text{---} \bullet \end{array} \quad (31)$$

$$\begin{array}{c} X \otimes Y \text{---} \bullet \begin{array}{l} \nearrow \quad \searrow \\ \nearrow \quad \searrow \end{array} \quad X \otimes Y \text{---} \bullet \begin{array}{l} \nearrow \quad \searrow \\ \nearrow \quad \searrow \end{array} \end{array} = \begin{array}{c} X \text{---} \bullet \begin{array}{l} \nearrow \quad \searrow \\ \nearrow \quad \searrow \end{array} \quad X \text{---} \bullet \begin{array}{l} \nearrow \quad \searrow \\ \nearrow \quad \searrow \end{array} \\ Y \text{---} \bullet \begin{array}{l} \nearrow \quad \searrow \\ \nearrow \quad \searrow \end{array} \quad Y \text{---} \bullet \begin{array}{l} \nearrow \quad \searrow \\ \nearrow \quad \searrow \end{array} \end{array} \quad (32)$$

and the naturality of del , which means that

$$\begin{array}{c} \text{---} \boxed{f} \text{---} \bullet \quad \text{---} \bullet \\ \text{---} \bullet \quad \text{---} \bullet \end{array} = \begin{array}{c} \text{---} \bullet \\ \text{---} \bullet \end{array} \quad (33)$$

5.3 Iterated copy maps and plates

The previous definitions are standard for Markov categories. We extend the graphical notation with n -fold maps and plates, which stand for tensor products repeated n times.

Definition 5.4 (*n*-fold copy map). The *n*-fold copy map $\text{copy}_X^n : X \rightarrow X^n$ is given by

$$\text{copy}_X^1 = \text{copy}_X \quad (34)$$

$$\text{copy}_X^n = \begin{array}{c} \boxed{\text{copy}_X^{n-1}} \\ \text{---} \bullet \text{---} \end{array} \quad n > 1 \quad (35)$$

In a string diagram, a plate that is annotated $i \in A$ means the tensor product of the $|A|$ elements that appear inside the plate. A wire crossing from outside a plate boundary to the inside of a plate indicates an $|A|$ -fold copy map, which we indicate by placing a dot on the plate boundary. We do not define anything that allows wires to cross from the inside of a plate to the outside; wires must terminate within the plate.

Thus, given $\mathbb{K}_i : X \rightarrow Y$ for $i \in A$,

$$\bigotimes_{i \in A} \mathbb{K}_i := \boxed{\begin{array}{c} \boxed{\mathbb{K}_i} \\ i \in A \end{array}} \text{copy}_X^{|A|} \left(\bigotimes_{i \in A} \mathbb{K}_i \right) := \text{---} \bullet \boxed{\begin{array}{c} \boxed{\mathbb{K}_i} \\ i \in A \end{array}} \quad (36)$$

5.3.1 Examples

String diagrams can always be converted into definitions involving integrals and tensor products. A number of shortcuts can help to make the translations efficiently.

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

$$\begin{array}{c} \text{---} \boxed{\mathbb{L}} \text{---} \bullet \boxed{\mathbb{K}} \text{---} \\ \text{---} \end{array} = (\text{id}_X \otimes \mathbb{L})\mathbb{K} \quad (37)$$

$$[(\text{id}_X \otimes \mathbb{L})\mathbb{K}](A|x, w) = \int_Y \int_X \mathbb{K}(A|x', y') \mathbb{L}(dy'|w) \delta_x(dx') \quad (38)$$

$$= \int_Y \mathbb{K}(A|x, y') \mathbb{L}(dy'|w) \quad (39)$$

That is, an identity map “passes its input directly to the next kernel”.

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

$$\begin{array}{c} \text{---} \bullet \text{---} \boxed{\mathbb{K}} \text{---} \\ \text{---} \end{array} = (\text{id}_X \otimes \text{copy}_Y)\mathbb{K} \quad (40)$$

$$[(\text{id}_X \otimes \text{copy}_Y)\mathbb{K}](A|x, y) = \int_Y \int_Y \mathbb{K}(A|x, y', y'') \delta_y(dy') \delta_y(dy'') \quad (41)$$

$$= \mathbb{K}(A|x, y, y) \quad (42)$$

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

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

$$\begin{array}{c} \text{---} \times \text{---} \text{---} \boxed{\mathbb{K}} \text{---} \\ \text{---} \end{array} = \text{swap}_{YX} \mathbb{K} \quad (43)$$

$$(\text{swap}_{YX} \mathbb{K})(A|y, x) = \int_{X \times Y} \mathbb{K}(A|x', y') \delta_y(dy') \delta_x(dx') \quad (44)$$

$$= \mathbb{K}(A|x, y) \quad (45)$$

The swap map before a kernel switches the input arguments.

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

$$\begin{array}{c} \text{---} \boxed{\mathbb{K}} \times \text{---} \\ \text{---} \end{array} = \mathbb{K} \text{swap}_{YZ} \quad (46)$$

$$(\mathbb{K} \text{swap}_{YZ})(A \times B|x) = \int_{Y \times Z} \delta_y(B) \delta_z(A) \mathbb{K}(dy \times dz|x) \quad (47)$$

$$= \int_{B \times A} \mathbb{K}(dy \times dz|x) \quad (48)$$

$$= \mathbb{K}(B \times A|x) \quad (49)$$

6 Probability sets

A probability set is a set of probability measures. This section establishes a number of useful properties of conditional probability with respect to probability sets. Unlike conditional probability with 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 6.1 (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 4.8 and 4.13 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 6.7) with respect to probability sets.

Definition 6.2 (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 $\mathbb{P}_{\{\}}^X$, 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 6.3 (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 6.4 (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 9.4). $\mathbb{P}_{\{\}}^{Z|(Y|X)}$ is in turn a version of the uniform conditional distribution $\mathbb{P}_{\{\}}^{Z|YX}$ (Theorem 9.5). Thus, from the existence of $\mathbb{P}_{\{\}}^{ZY|X}$ we can derive the existence of $\mathbb{P}_{\{\}}^{Z|YX}$.

6.1 Semidirect product and almost sure equality

The operation used in Equation 13 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 6.5 (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

$$\mathbb{K} \odot \mathbb{L} := \text{copy}_X(\mathbb{K} \otimes \text{id}_X)(\text{copy}_Y \otimes \text{id}_X)(\text{id}_Y \otimes \mathbb{L}) \quad (50)$$

(51)

$$\Longleftrightarrow \tag{52}$$

$$(\mathbb{K} \odot \mathbb{L})(A \times B|x) = \int_A \mathbb{L}(B|y, x) \mathbb{K}(dy|x) \quad A \in \mathcal{Y}, B \in \mathcal{Z} \quad (53)$$

Lemma 6.6 (Semidirect product is associative). *Given $\mathbb{K} : X \rightarrow Y$, $\mathbb{L} : Y \times X \rightarrow Z$ and $\mathbb{M} : Z \times Y \times X \rightarrow W$*

$$(\mathbf{K} \odot \mathbf{L}) \odot \mathbf{Z} = \mathbf{K} \odot (\mathbf{L} \odot \mathbf{Z}) \quad (54)$$

(55)

Proof.

$$(\mathbb{K} \odot \mathbb{L}) \odot \mathbb{M} = \begin{array}{c} \text{Diagram showing the composition of three maps } \mathbb{K}, \mathbb{L}, \text{ and } \mathbb{M} \text{ on a 4-strand braid.} \\ \text{Input strands: } X, Y, W, Z \text{ (from top to bottom).} \\ \text{Map } \mathbb{K} \text{ (top-left box):} \\ \quad X \rightarrow Y \\ \quad Y \rightarrow X \\ \quad W \rightarrow W \\ \quad Z \rightarrow Z \\ \text{Map } \mathbb{L} \text{ (top-right box):} \\ \quad X \rightarrow W \\ \quad Y \rightarrow Z \\ \quad W \rightarrow Y \\ \quad Z \rightarrow X \\ \text{Map } \mathbb{M} \text{ (bottom box):} \\ \quad X \rightarrow X \\ \quad Y \rightarrow Y \\ \quad W \rightarrow Z \\ \quad Z \rightarrow W \end{array} \quad (56)$$

$$= \begin{array}{c} \text{---} X \\ \text{---} Y \\ \text{---} X \\ \text{---} Y \\ \text{---} W \\ \text{---} Z \end{array} \quad (57)$$

$$= \mathbb{K} \odot (\mathbb{L} \odot \mathbb{M}) \quad (58)$$

☐

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

Definition 6.7 (Almost sure equality). Two Markov kernels $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \rightarrow Y$ are almost surely equal $\stackrel{\mathbb{P}_{\Omega}}{\cong}$ with respect to a probability set \mathbb{P}_{Ω} and variable $X : \Omega \rightarrow X$ if for all $\mathbb{P}_{\alpha} \in \mathbb{P}_{\Omega}$,

$$\mathbb{P}_\alpha^X \odot K = \mathbb{P}_\alpha^X \odot L \quad (59)$$

Lemma 6.8 (Uniform conditional distributions are almost surely equal). *If $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \rightarrow Y$ are both versions of $\mathbb{P}_{\{\cdot\}}^{Y|X}$ then $\mathbb{K} \stackrel{\mathbb{P}_0}{\cong} \mathbb{L}$*

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

$$\mathbb{P}_\alpha^{\mathbf{X}} \odot \mathbb{K} = \mathbb{P}_\alpha^{\mathbf{XY}} \quad (60)$$

$$= \mathbb{P}_\alpha^{\mathbf{X}} \odot \mathbb{L} \quad (61)$$

□

Lemma 6.9 (Substitution of almost surely equal Markov kernels). *Given $\mathbb{P}_\{\}$, if $\mathbb{K} : X \times Y \rightarrow Z$ and $\mathbb{L} : X \times Y \rightarrow Z$ are almost surely equal $\mathbb{K} \stackrel{\mathbb{P}_\{\}}{\cong} \mathbb{L}$, then for any $\mathbb{P}_\alpha \in \mathbb{P}_\{\}$*

$$\mathbb{P}_\alpha^{\mathbf{Y|X}} \odot \mathbb{K} \stackrel{\mathbb{P}_\{\}}{\cong} \mathbb{P}_\alpha^{\mathbf{Y|X}} \odot \mathbb{L} \quad (62)$$

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

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

$$= \mathbb{P}_\alpha^{\mathbf{X}} \odot (\mathbb{P}_\{\}^{\mathbf{Y|X}} \odot \mathbb{K}) \quad (64)$$

$$= \mathbb{P}_\alpha^{\mathbf{X}} \odot (\mathbb{P}_\{\}^{\mathbf{Y|X}} \odot \mathbb{L}) \quad (65)$$

□

Theorem 6.10 (Semidirect product of uniform conditional distributions is a joint uniform conditional distribution). *Given a probability set $\mathbb{P}_\{\}$ on (Ω, \mathcal{F}) , variables $\mathbf{X} : \Omega \rightarrow X$, $\mathbf{Y} : \Omega \rightarrow Y$ and uniform conditional distributions $\mathbb{P}_\{\}^{\mathbf{Y|X}}$ and $\mathbb{P}_\{\}^{\mathbf{Z|XY}}$, then $\mathbb{P}_\{\}^{\mathbf{YZ|X}}$ exists and is equal to*

$$\mathbb{P}_\{\}^{\mathbf{YZ|X}} = \mathbb{P}_\{\}^{\mathbf{Y|X}} \odot \mathbb{P}_\{\}^{\mathbf{Z|XY}} \quad (66)$$

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

$$\mathbb{P}_\alpha^{\mathbf{XYZ}} = \mathbb{P}_\alpha^{\mathbf{X}} \odot \mathbb{P}_\alpha^{\mathbf{YZ|X}} \quad (67)$$

$$= \mathbb{P}_\alpha^{\mathbf{X}} \odot (\mathbb{P}_\alpha^{\mathbf{Y|X}} \odot \mathbb{P}_\alpha^{\mathbf{Z|YX}}) \quad (68)$$

$$= \mathbb{P}_\alpha^{\mathbf{X}} \odot (\mathbb{P}_\{\}^{\mathbf{Y|X}} \odot \mathbb{P}_\{\}^{\mathbf{Z|YX}}) \quad (69)$$

□

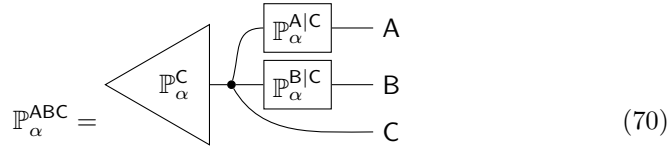
6.2 Conditional independence

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 6.11) or in terms of the existence of a conditional distribution that ignores one of its inputs (Theorem 6.12).

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 6.14). We will discuss in Section 3 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 6.11 (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

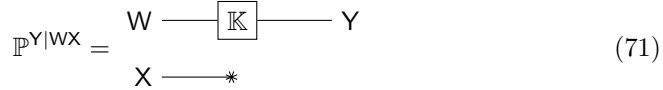


$$\mathbb{P}_\alpha^{ABC} = \begin{array}{c} \text{triangle} \\ \mathbb{P}_\alpha^C \end{array} \begin{array}{l} \text{box } \mathbb{P}_\alpha^{A|C} \text{ --- } A \\ \text{box } \mathbb{P}_\alpha^{B|C} \text{ --- } B \\ \text{--- } C \end{array} \quad (70)$$

Cho and Jacobs (2019) have shown that this definition coincides with the standard notion of conditional independence for a particular probability model (Theorem 6.12).

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

Theorem 6.12. *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*



$$\mathbb{P}^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (71)$$

$$\iff \mathbb{P}^{Y|WX}(A|w, x) = \mathbb{K}(A|w) \quad \forall A \in \mathcal{Y} \quad (72)$$

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

Theorem 6.13 shows how, under some circumstances, it is possible to infer an extended conditional independence in a probability set \mathbb{P}_C from a regular conditional independence that holds in one element of the set \mathbb{P}_α . We ultimately carry out the procedure associated with on only one element of C , so usually we cannot test whether some property holds for the whole set \mathbb{P}_C . However, regular conditional independences with respect to a particular element of \mathbb{P}_C can be tested for (again, subject to some assumptions (Shah and Peters, 2020)).

Theorem 6.13. *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

$$\mathbb{P}_C^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (73)$$

Proof. See Appendix 9.4 □

Corollary 6.14. *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 $\mathbb{P}_C^{Y|WX}$ 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 6.13, there is $\mathbb{K} : W \rightarrow Y$ such that for all β

$$\mathbb{P}_\beta^{WY} = \begin{array}{c} \begin{array}{c} \triangleleft \mathbb{P}_\alpha^{WX} \end{array} \begin{array}{c} \text{---} \bullet \text{---} W \\ \text{---} \bullet \text{---} * \\ \text{---} \bullet \text{---} \boxed{\mathbb{P}_C^{Y|WX}} \text{---} Y \end{array} \end{array} \quad (74)$$

$$= \begin{array}{c} \begin{array}{c} \triangleleft \mathbb{P}_\alpha^{WX} \end{array} \begin{array}{c} \text{---} \bullet \text{---} W \\ \text{---} \bullet \text{---} * \\ \text{---} \bullet \text{---} \boxed{\mathbb{K}} \text{---} Y \end{array} \end{array} \quad (75)$$

$$= \begin{array}{c} \begin{array}{c} \triangleleft \mathbb{P}_\alpha^{W} \end{array} \begin{array}{c} \text{---} \bullet \text{---} \boxed{\mathbb{K}} \text{---} Y \end{array} \end{array} \quad (76)$$

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

6.3 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 6.15 (Uniform conditional independence). Given a probability set \mathbb{P}_C and variables X , Y and Z , the uniform conditional independence $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e XC|Z$ holds if $\mathbb{P}_C^{Y|XZ}$ and $\mathbb{P}_C^{Y|X}$ exist and

$$\begin{array}{ccc} & Z & \text{---} \boxed{\mathbb{P}_C^{Y|Z}} \text{---} Y \\ \mathbb{P}_C^{Y|XZ} \stackrel{\mathbb{P}_C}{\cong} & X & \text{---} * \end{array} \quad (77)$$

$$\iff \quad (78)$$

$$\mathbb{P}_C^{Y|XZ}(A|x, z) \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y|Z}(A|z) \quad \forall A \in \mathcal{Y}, (x, z) \in X \times Z \quad (79)$$

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. 79, while we require that these functions are Markov kernels (they are measurable and probability distribution-value, as in Definition 4.3).

Example 6.16 (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 $A \in \mathcal{Y}$, $B \in \mathcal{C}$:

$$\mathbb{P}_\alpha^{YC}(A \times B) = \mathbb{P}_\alpha(A) \delta_\alpha(B) \quad (80)$$

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

$$\mathbb{P}_\alpha^{Y|D} = \mathbb{P}_\alpha^Y \quad (81)$$

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

$$\mathbb{P}_C^{Y|D}(A|\alpha) = \mathbb{P}_\alpha^Y(A) \quad \forall A \in \mathcal{Y}, \alpha \in C \quad (82)$$

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 6.17 (Uniform conditional independence representation). *Given a probability set \mathbb{P}_C with a uniform conditional probability $\mathbb{P}_C^{XY|Z}$,*

$$\mathbb{P}_C^{XY|Z} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \quad (83)$$

$$\iff \quad (84)$$

$$\mathbb{P}_C^{XY|Z}(A \times B|z) \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{X|Z}(A|z)\mathbb{P}_C^{Y|Z}(B|z) \quad \forall A \in \mathcal{X}, B \in \mathcal{Y}, z \in Z \quad (85)$$

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

Proof. If: By Theorem 9.5

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

$$\stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } * \text{ --- } X \end{cases} \end{array} \quad (87)$$

$$= \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \quad (88)$$

Only if: Suppose

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

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

$$\mathbb{P}_C^{Y|XZ}(B|x, z) > \mathbb{P}_C^{Y|Z}(B|z) \quad \forall (x, z) \in A \times C \quad (90)$$

then

$$\mathbb{P}_\alpha^{\text{XYZZ}}(A \times B \times C) = \int_{A \times C} \mathbb{P}_C^{\text{Y|XZ}}(B|x, z) \mathbb{P}_C^{\text{X|Z}}(\text{d}x|z) \mathbb{P}_\alpha^{\text{Z}}(\text{d}z) \quad (91)$$

$$> \int_{A \times C} \mathbb{P}_C^{\text{Y|X}}(B|z) \mathbb{P}_C^{\text{X|Z}}(\text{d}x|z) \mathbb{P}_\alpha^{\text{Z}}(\text{d}z) \quad (92)$$

$$= \int_C \mathbb{P}_C^{\text{XY|X}}(A \times B|z) \mathbb{P}_\alpha^{\text{Z}}(\text{d}z) \quad (93)$$

$$= \mathbb{P}_\alpha^{\text{XYZZ}}(A \times B \times C) \quad (94)$$

a contradiction. An analogous argument follows if we replace “>” with “<” in Eq. 90. \square

7 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 $\mathbb{P}_C^{\text{Y|X}}$ exists, then the model says that the joint outcome $\mathbb{P}_\alpha^{\text{XY}}$ of any choice $\alpha \in C$ can be computed from the marginal distribution $\mathbb{P}_\alpha^{\text{X}}$ alone. We are going to ask the question: in which kind of probability sets do uniform conditionals of the form $\mathbb{P}_C^{\text{Y|XH}}$ exist? Here H is a “fixed but unknown” hypothesis that becomes better known as more data is observed. Roughly speaking, $\mathbb{P}_C^{\text{Y|XH}}$ represents the response of Y to X regardless of which choice is made.

A decision makers may be interested in a functions like $\mathbb{P}_C^{\text{Y|XH}}$. 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 $\mathbb{P}_C^{\text{Y|XH}}$ a *response conditional*.

If a model \mathbb{P}_C supports $\mathbb{P}_C^{\text{Y|XH}}$, 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 $\mathbb{P}_C^{\text{Y|XH}}$ 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 $\mathbb{P}_C^{\text{Y|XH}}$. Secondly, we consider what kind of decision procedures can be modeled by probability sets of this type.

7.1 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 $\mathbf{Y} := (\mathbf{Y}_i)_{i \in M}$ (the outcome sequence) and $\mathbf{D} := (\mathbf{D}_i)_{i \in M}$ (the action sequence) for some index set $M \subset \mathbb{N}$. We say \mathbf{Y}_i corresponds to \mathbf{D}_i . We are specifically looking for uniform conditionals of the form $\mathbb{P}_C^{\mathbf{Y}_i | \mathbf{D}_i \mathbf{H}}$ for all $i \in M$. \mathbf{H} here is a hypothesis, and it must be fixed with respect to different choices – i.e. $\mathbf{H} \perp\!\!\!\perp_{\mathbb{P}_C}^e C$.

We assume a starting point that $\mathbb{P}_C^{\mathbf{Y} | \mathbf{D}}$ exists. This could be guaranteed, for example, if each choice α corresponds to a unique deterministic distribution $\mathbb{P}_\alpha^{\mathbf{D}}$ (see Example 6.16).

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 $\mathbb{P}_C^{\mathbf{Y} | \mathbf{D}}$ 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 $A \subset M$, $\mathbb{P}_C^{\mathbf{Y}_A | \mathbf{D}_A}$ exists and

$$\mathbb{P}_C^{\mathbf{Y}_A | \mathbf{D}_M} = \mathbf{D}_{M \setminus A} \longrightarrow \left[\mathbb{P}_C^{\mathbf{Y}_A | \mathbf{D}_A} \right] \mathbf{Y}_A \quad (95)$$

In the language of extended conditional independence, it is the assumption $\mathbf{Y}_A \perp\!\!\!\perp_{\mathbb{P}_C}^e C \mathbf{D}_{M \setminus A} | \mathbf{D}_A$.

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 \mathbf{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):

(96)

Exchange commutativity (swapping choices \sim swapping labels):

(97)

—end not sure where to put—

7.2 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 7.1 (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 7.2 (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 $\rho : M \rightarrow M$

$$\text{swap}_{\rho(\mathbf{D})} \mathbb{P}_C^{\mathbf{Y}|\mathbf{D}} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{\mathbf{Y}|\mathbf{D}} \text{swap}_{\rho(\mathbf{Y})} \quad (98)$$

Then $\mathbb{P}_C^{\mathbf{Y}|\mathbf{D}}$ is $(\mathbf{D}; \mathbf{Y})$ -exchange commutative.

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

Definition 7.3 (Locality of consequences). Suppose we have a sample space (Ω, \mathcal{F}) and a probability set \mathbb{P}_C with uniform conditional probability $\mathbb{P}_C^{\mathbf{Y}|\mathbf{D}}$ 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 $A \subset M$

$$\mathbb{P}_S^{\mathbf{Y}_A|\mathbf{D}_M} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} \mathbf{D}_A \text{ --- } \boxed{\mathbb{P}_C^{\mathbf{Y}_A|\mathbf{D}_A}} \text{ --- } \mathbf{Y}_A \\ \mathbf{D}_{M \setminus A} \text{ --- } * \end{array} \quad (99)$$

then $\mathbb{P}_C^{\mathbf{Y}|\mathbf{D}}$ exhibits $(\mathbf{D}; \mathbf{Y})$ -local consequences.

If \mathbb{P}_C exhibits $(\mathbf{D}; \mathbf{Y})$ -local consequences then, given two different choices α and α' such that $\mathbb{P}_\alpha^{\mathbf{D}_A} = \mathbb{P}_{\alpha'}^{\mathbf{D}_A}$ then $\mathbb{P}_\alpha^{\mathbf{Y}_A} = \mathbb{P}_{\alpha'}^{\mathbf{Y}_A}$. However, \mathbb{P}_C may exhibit consequence locality even if no such pair of choices exists.

Theorem 7.4 shows that neither condition implies the other.

Theorem 7.4. *Exchange commutativity does not imply locality of consequences or vice versa.*

Proof. Appendix 9.6. □

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 $\mathbf{D}_1 = 0$), 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 $\mathbf{D}_1 = 1$). At the second step, I may further choose to copy the provisional results ($\mathbf{D}_2 = 0$) or invert them ($\mathbf{D}_2 = 1$). Then

$$\mathbb{P}_S^{Y_1|D}(y_1|d_1, d_2) = 0.5 \quad (100)$$

$$\mathbb{P}_S^{Y_2|D}(y_2|d_1, d_2) = 0.5 \quad (101)$$

- The marginal distribution of both experiments in isolation is Bernoulli(0.5) no matter what choices I make, so a model of this experiment would satisfies Definition 7.3
- Nevertheless, the choice for the first experiment affects the result of the second experiment

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

Definition 7.5 (Causal contractibility). A probability set \mathbb{P}_C with uniform conditional $\mathbb{P}_C^{Y|D}$ is $(D; Y)$ -*causally contractible* if it is both exchange commutative and exhibits consequence locality.

Theorem 7.6 (Equality of conditionals). *A probability set \mathbb{P}_C that is $(D; Y)$ -causally contractible has, for any $A, B \subset M$ with $|A| = |B|$*

$$\mathbb{P}_C^{Y_A|D_A} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_B|D_B} \quad (102)$$

Proof. 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.

$$\begin{array}{c} D_A \text{ --- } \boxed{\mathbb{P}_C^{Y_A|D_A}} \text{ --- } Y_A \\ D_{M \setminus A} \text{ --- } * \end{array} = \begin{array}{c} D_{M \setminus A} \text{ --- } \boxed{\mathbb{P}_C^{Y_A Y_{M \setminus A} | D_A D_{M \setminus A}}} \text{ --- } Y_A \\ D_{M \setminus A} \text{ --- } * \end{array} \quad (103)$$

$$= \begin{array}{c} D_{M \setminus A} \text{ --- } \boxed{s_{BA}} \text{ --- } \boxed{\mathbb{P}_C^{Y_A Y_{M \setminus A} | D_A D_{M \setminus A}}} \text{ --- } \boxed{s_{AB}} \text{ --- } Y_A \\ D_{M \setminus A} \text{ --- } * \end{array} \quad (104)$$

$$= \begin{array}{c} D_{M \setminus B} \text{ --- } \boxed{\mathbb{P}_C^{Y_B Y_{M \setminus B} | D_A D_{M \setminus B}}} \text{ --- } Y_B \\ D_{M \setminus B} \text{ --- } * \end{array} \quad (105)$$

Thus

$$\mathbb{P}_C^{Y_A | D_A D_{M \setminus A}} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_B | D_B D_{M \setminus B}} \quad (106)$$

$$\stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} D_A \text{ --- } \boxed{\mathbb{P}_C^{Y_A | D_A}} \text{ --- } Y_A \\ D_{M \setminus A} \text{ --- } * \end{array} \quad (107)$$

$$\implies \mathbb{P}_C^{Y_A | D_A} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_B | D_B} \quad (108)$$

□

7.3 Existence of response conditionals

The main result in this section is Theorem 7.9 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 $\mathbb{P}_C^{Y_1|D_1H}$. 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.

Theorem 7.7. *Given a probability set \mathbb{P}_C such that $D := (D_i)_{i \in \mathbb{N}}$ and $Y := (Y_i)_{i \in \mathbb{N}}$, \mathbb{P}_C is $(D; Y)$ -causally contractible if and only if there exists a column exchangeable probability distribution $\mu^{Y^D} \in \Delta(Y^{|D| \times \mathbb{N}})$ such that*

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \triangle \\ \mu^{Y^D} \\ D \text{ --- } \square \text{F}_{ev} \text{ --- } Y \end{array} \quad (109)$$

$$\iff \quad (110)$$

$$\mathbb{P}_C^{Y|D}(y|(d_i)_{i \in \mathbb{N}}) = \mu^{Y^D} \Pi_{(d_i i)_{i \in \mathbb{N}}}(y) \quad (111)$$

Where $\Pi_{(d_i i)_{i \in \mathbb{N}}} : Y^{|D| \times \mathbb{N}} \rightarrow Y^{\mathbb{N}}$ 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

$$ev : D^{\mathbb{N}} \times Y^{D \times \mathbb{N}} \rightarrow Y \quad (112)$$

$$((d_i)_{i \in \mathbb{N}}, (y_{ij})_{i \in D, j \in \mathbb{N}}) \mapsto (y_{d_i i})_{i \in \mathbb{N}} \quad (113)$$

Proof. Appendix 9.6. □

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 109. Thus we define an augmented causally contractible model as follows:

this feels really kludgy, but not having it is also a pain

Definition 7.8 (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 $Y^D : \Omega \rightarrow Y^{|D| \times \mathbb{N}}$ such that $\mathbb{P}_C^{Y^D}$ exists and

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \triangle \\ \mathbb{P}_C^{Y^D} \\ D \text{ --- } \square \text{F}_{ev} \text{ --- } Y \end{array} \quad (114)$$

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 7.7 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_i^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 $\mathbb{P}_C^{Y^D|H}$ 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 $\mathbb{P}_C^{Y_i|D_iH}$ and the same common prior \mathbb{P}_C^H .

Theorem 7.9. *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*

$$\mathbb{P}_C^{Y^D} = \begin{array}{c} \begin{array}{c} \triangle \mu^H \\ \text{---} \end{array} \begin{array}{c} \text{---} \\ \text{---} \end{array} \begin{array}{c} \boxed{\Pi_{D,i}} \end{array} \begin{array}{c} \boxed{\mathbb{P}_{\square}^{Y_0|HD_0}} \end{array} \text{---} Y_i \\ \text{---} D \end{array} \quad i \in \mathbb{N} \quad (115)$$

$$\iff \quad (116)$$

$$Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e Y_{N \setminus i}, D_{N \setminus i} C | HD_i \quad \forall i \in \mathbb{N} \quad (117)$$

$$\wedge H \perp\!\!\!\perp_{\mathbb{P}_C}^e DC \quad (118)$$

$$\wedge \mathbb{P}_C^{Y_i|HD_i} = \mathbb{P}^{Y_0|HD_0} \quad \forall i \in \mathbb{N} \quad (119)$$

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

Proof. Appendix 9.6. □

7.4 Example: backdoor adjustment

Suppose a probability set \mathbb{P}_C is $(D, X; Y)$ -causally contractible and $X_i \perp\!\!\!\perp_{\mathbb{P}_C}^e D_i C | H$. Then

$$\mathbb{P}_\alpha^{Y_i | D_i H}(A | d, h) = \int_X \mathbb{P}_\alpha^{Y_i | X_i D_i H}(A | d, x, h) \mathbb{P}_\alpha^{X_i | D_i H}(dx | d, h) \quad (120)$$

$$= \int_X \mathbb{P}_C^{Y_i | X_i D_i H}(A | d, x, h) \mathbb{P}_C^{X_i | H}(dx | h) \quad (121)$$

If we additionally assume $\mathbb{P}_C^{X_i | H} \cong \mathbb{P}_C^{X_1 | H}$ then

$$\mathbb{P}_\alpha^{Y_i | D_i H}(A | d, h) = \int_X \mathbb{P}_C^{Y_i | X_i D_i H}(A | d, x, h) \mathbb{P}_C^{X_1 | H}(dx | h) \quad (122)$$

Equation 122 is structurally identical to the backdoor adjustment formula for an intervention on D_1 targeting Y_1 where X_1 is a common cause of both. Note that we have not assumed that $\mathbb{P}_C^{D_i}$ exists for any i ; the only assumptions are the extended independence $X_i \perp\!\!\!\perp_{\mathbb{P}_C}^e D_i C | H$ and the equality in distribution between the first and i th “ X ”.

7.5 Assessing causal contractibility

We have a formal condition – causal contractibility – equivalent to the existence of response conditionals. A challenge we still need to face is to evaluate when a decision procedure is appropriately modeled with a probability set causally contractible with respect to some pair of variables.

Similar questions have been examined before, and a variety of answers have been given. These answers all say something like: the experiment consists of a sequence of similar or interchangeable “individuals” or “units”, and that “treatment assignments” must be independent of the individual’s identities, or may be swapped without swapping the individuals. For example, Greenland and Robins (1986) explain

Equivalence of response type may be thought of in terms of exchangeability of individuals: if the exposure states of the two individuals had been exchanged, the same data distribution would have resulted.

Dawid (2021):

A group of individuals whom I can regard, in an intuitive sense, as similar to myself, with headaches similar to my own. [...] Whether or not the exchangeability assumption can be regarded as reasonable will be highly dependent on the background information informing my personal probability assessments

Rubin (2005):

indexing of the units is, by definition, a random permutation of $1, \dots, N$, and thus any distribution on the science must be row-exchangeable [...]. The second critical fact is that if the treatment assignment mechanism is ignorable (e.g., randomized), then when the expression for the assignment mechanism (2) is evaluated at the observed data, it is free of dependence on Y_{mis}

Theorem 7.10 formalises these ideas. As an example of its application, consider an experiment where N patients are treated, each with an individual identifier I_i , who receives treatment X_i and experiences outcome Y_i . We assume a $(X; I; Y)$ -causally contractible model is appropriate. This reflects two judgments; firstly, that treatments and identifiers only affect their local consequences (Definition 7.3), and secondly we are indifferent the order in which the individuals appear and the treatments are received. We also assume that at the outset we are ignorant about any differences between each individual; specifically, permuting the individuals involved in the experiment leaves us with the same model.

Under these conditions, additionally assuming that treatments are independent of identifiers conditional on the hypothesis H implies that the model is also $(X; Y)$ -causally contractible. This last assumption is somewhat difficult to interpret. To understand its meaning, consider that each identifier $a \in I$ is associated, in the large sample limit $h \in H$, with a stochastic “individual response function” $\mathbb{K} : X \rightarrow Y$ defined by

$$\mathbb{K}_{a,h}(A|x) := \mathbb{P}_C^{Y_i|X_i I_i H}(A|x, a, h) \quad (123)$$

Then the assumption $I_i \perp\!\!\!\perp_{\mathbb{P}_C}^e X_i C | H$ means that X_i is uninformative about the individual response function $\mathbb{K}_{I_i, H}$ (abusing notation). This is strongly reminiscent of the “ignorability” assumption found in the Potential Outcomes literature (Rubin, 2005).

Theorem 7.10 (Causal contractibility with independent treatments). *Suppose we have a probability set \mathbb{P}_C , $(X; I; Y)$ -causally contractible for $\mathbf{X} := (X_i)_{i \in \mathbb{N}}$, $\mathbf{I} := (I_i)_{i \in \mathbb{N}}$ and $\mathbf{Y} := (Y_i)_{i \in \mathbb{N}}$, $I_i : \Omega \rightarrow I$ countable and $X_i : \Omega \rightarrow X$, $Y_i : \Omega \rightarrow Y$ standard measurable.*

For arbitrary finite permutation $\rho : I \rightarrow I$, let $\mathbb{F}_{\rho_I} : X^{\mathbb{N}} \times I^{\mathbb{N}} \rightarrow X^{\mathbb{N}} \times I^{\mathbb{N}}$ be the Markov kernel associated with the function

$$\rho_I : (x_i, n_i)_{i \in \mathbb{N}} \mapsto (x_i, \rho(n_i))_{i \in \mathbb{N}} \quad (124)$$

Suppose for any ρ

$$\mathbb{P}_C^{Y|XI} \stackrel{\mathbb{P}_C}{\cong} \mathbb{F}_{\rho_I} \mathbb{P}_C^{Y|XI} \quad (125)$$

and for all $i \in \mathbb{N}$, $I_i \perp\!\!\!\perp_{\mathbb{P}_C}^e X_i C | H$.

Then \mathbb{P}_C is $(X; Y)$ -causally contractible.

Proof. By causal contractibility, we have for all $i \in |I|$

$$\mathbb{P}_C^{Y_i|X_i H_{|I|}} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_1|X_1 H_1} \otimes \text{erase}_{I \setminus \{i\}} \quad (126)$$

Furthermore, by assumption, for arbitrary permutation $\rho : I \rightarrow I$

$$\mathbb{P}_C^{Y_i|X_i H_i}(A|x, h, n) \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_i|X_i H_i}(A|x, h, n, m) \quad \forall m \in I \quad (127)$$

$$\stackrel{\mathbb{P}_C}{\cong} (\mathbb{F}_{\rho_I} \mathbb{P}_C^{Y_i|X_i H_{|I|}})(A|x, h, n, m) \quad (128)$$

$$\stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_i|X_i H_{|I|}}(A|x, h, \rho(n, m)) \quad (129)$$

$$\stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_1|X_1 H_1}(A|x, h, \Pi_i(\rho(n, m))) \quad (130)$$

Where Π_i projects the i -th coordinate of $\rho(n, m)$. Because m and ρ are arbitrary, this implies for any $p \in I$

$$\mathbb{P}_C^{Y_i|X_i H_i}(A|x, h, n) \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_1|X_1 H_1}(A|x, h, p) \quad (131)$$

$$\implies \mathbb{P}_C^{Y_i|X_i H_i} \stackrel{\mathbb{P}_C}{\cong} \mathbb{K} \otimes \text{erase}_I \quad (132)$$

where $\mathbb{K} : (A|x, h) \mapsto \mathbb{P}_C^{Y_i|X_i H_i}(A|x, h, q)$ for some $q \in I$. Therefore $Y_i \perp_{\mathbb{P}_C}^e I_i C | X_i H$.

Furthermore, by assumption, $I_i \perp_{\mathbb{P}_C}^e X_i C | H$, and so by weak union $Y_i \perp_{\mathbb{P}_C}^e I_i C | X_i H$ and \mathbb{P}_C is therefore $(X; Y)$ -causally contractible by Lemma 9.20. \square

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 X_M corresponds to \mathcal{X}_M and so forth. If \mathbb{P}_C is (X_{2M}, Y_{2M}) -causally contractible model then the following holds (see corollary 7.6):

$$\mathbb{P}_C^{Y_{[2, M+1]} | X_{[2, M+1]}} = \mathbb{P}^{Y_{(M, 2M)} | X_{(M, 2M)}} \quad (133)$$

$$\implies \mathbb{P}_C^{Y_{M+1} | X_{[2, M+1]} Y_{[2, M]}} = \mathbb{P}^{Y_{M+1} | X_{(M, 2M)} Y_{(M+1, 2M)}} \quad (134)$$

That is, causal contractibility implies that there is no difference between conditioning on observational results or on the results of active choices; active choices are as good for predicting observations as vice-versa.

When can such a strong assumption be accepted? Like exchangeability, one option for justifying judgements of causal contractibility is to compare different measurement procedures and argue that they should be *indistinguishable* – that is, they should be modeled with exactly the same model. In the case of exchangeability, this is sometimes called “symmetric ignorance” Gelman et al. (2021, chap. 5). More specifically, if we are committed to modelling uncertainty with a single probability distribution and we consider the following two measurement procedures indistinguishable:

- 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.

We can make a similar argument for causal contractibility, although it is a substantially more complicated one. We will focus only on commutativity of exchange here. There is some discussion in literature of procedural assumptions that imply properties similar to commutativity of exchange. For example,

7.6 Body mass index revisited

If we have a probability set \mathbb{P}_C with $\mathbf{B} := (\mathbf{B}_i)_{i \in M}$ representing body mass index and $\mathbf{Y} := (\mathbf{Y}_i)_{i \in M}$ representing health outcomes of interest, the justifications for causal effects in the previous section do not apply. The reason is that we are interested in evaluating actions that do not set body mass deterministically, whether directly or on the basis of a random signal.

However, it is still possible that \mathbb{P}_C could be $(\mathbf{B}; \mathbf{Y})$ -causally contractible. The key is Theorem 6.13 – if we start with \mathbb{P}_C that is $(\mathbf{D}; \mathbf{Y})$ -causally contractible, and we find for a suitable α that $\mathbf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathbf{D}_i | \mathbf{H}\mathbf{B}_i$ for all i , then we can conclude that $\mathbf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{D}_i | \mathbf{H}\mathbf{B}_i$ for all i , and that \mathbb{P}_C is $(\mathbf{B}; \mathbf{Y})$ -causally contractible (Theorem 7.11). There are two important things to bear in mind:

- Causal contractibility is always relative to a set of choices C ; our theory provides no way to assess the existence of “causal effects” independent of such a set
- Whether or not $\mathbf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathbf{D}_i | \mathbf{H}\mathbf{B}_i$ holds for a suitable choice α is a testable question, so it may be inappropriate to assume it does by asking at the outset about “the causal effect of \mathbf{B} ”

Thus it is possible in principle to have a “causal effect of body mass index”, but our analysis suggests that it should be demonstrated rather than assumed.

Theorem 7.11. *Suppose we have a probability set \mathbb{P}_C that is $(\mathbf{D}; \mathbf{X}, \mathbf{Y})$ -causally contractible, where $\mathbf{D} := (\mathbf{D}_i)_{i \in M}$ and likewise for \mathbf{X} and \mathbf{Y} . If there exists $\alpha \in C$ such that $\mathbb{P}_\alpha^{\mathbf{D}} \gg \{\mathbb{P}_\beta^{\mathbf{D}} | \beta \in C\}$ and $\mathbf{Y}_i \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathbf{D}_i | \mathbf{H}\mathbf{X}_i$ for all $i \in M$, then \mathbb{P}_C is also $(\mathbf{Y}; \mathbf{X})$ -causally contractible.*

Proof. See Appendix 9.7 □

7.7 Weakening causal contractibility

We have pointed out that causal contractibility is a very strong assumption, and will usually be unacceptable. We proposed three assumptions that could justify causal contractibility for some decision procedures:

1. Two procedures \mathcal{S}_α and $\mathcal{S}_{\alpha'}$ are indistinguishable if the description of one can be obtained by permuting patients in the description of the other (*patient indistinguishability*)
2. Two procedures \mathcal{S}_α and $\mathcal{S}_{\alpha'}$ are indistinguishable if the description of one can be obtained by permuting the order of treatment administration in the description of the other (*order indistinguishability*)
3. $c : A \rightarrow D^2$ is an invertible function

We can imagine that many decision procedures involving patient treatment might satisfy the first two properties, but not the third. A question to explore is: is there a representation theorem relevant to a decision procedure in which the third assumption is removed or weakened? A common assumption in other causal frameworks for procedures like this is to assume that there is some variable (X, Z) that is “causally sufficient” for Y , but Z is not observed. In our framework, this would amount to \mathbb{P}_C being $((X, Z); Y)$ -causally contractible, with Z unobserved. However, at this point we do not know of a relevant representation theorem for unobservable causal contractibility like this, nor of an argument that connects the representation with a particular family of decision procedures.

8 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 thing causal

contractibility might hold and experiments where we think it might not. The differences between the two can sometimes be subtle. This refines the idea put forward by Hernán (2016) that potential outcomes are well-defined when they are suitably precisely specified; in particular, we argue that the necessary kind of “precision” is that actions are deterministically specified when the decision maker’s knowledge is consistent with a judgement of causal contractibility.

There are two challenges that arise when we try to apply this approach to typical causal inference problems. The first is that choice variables (that is, variables that represent a decision maker’s choices) play a prominent role in our theory but in many common causal investigations they do not play such a role. Strictly speaking, conditional probability models may be 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) .

8.1 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.

9 Appendix, needs to be organised

9.1 Markov categories

Fritz (2020) defines Markov categories in the following way:

Definition 9.1. A Markov category C is a symmetric monoidal category in which every object $X \in C$ is equipped with a commutative comonoid structure given by a comultiplication $\text{copy}_X : X \rightarrow X \otimes X$ and a counit $\text{del}_X : X \rightarrow I$, depicted in string diagrams as

$$\text{del}_X := \text{---} * \text{copy}_X \qquad \qquad \qquad := \text{---} \bullet \text{---} \quad (135)$$

and satisfying the commutative comonoid equations

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \\ \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \end{array} = \begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \\ \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \end{array} \quad (136)$$

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \\ \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \end{array} = \begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \\ \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \end{array} \quad (137)$$

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \\ \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \end{array} = \begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \\ \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \end{array} \quad (138)$$

as well as compatibility with the monoidal structure

$$\begin{array}{c} X \otimes Y \text{---} \bullet \\ X \otimes Y \text{---} \bullet \end{array} = \begin{array}{c} X \text{---} \bullet \\ X \text{---} \bullet \end{array} \quad (139)$$

$$\begin{array}{c} X \otimes Y \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \\ X \otimes Y \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \end{array} = \begin{array}{c} X \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \\ Y \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \end{array} \quad (140)$$

and the naturality of del , which means that

$$\begin{array}{c} \text{---} \boxed{f} \text{---} \bullet \\ \text{---} \boxed{f} \text{---} \bullet \end{array} = \begin{array}{c} \text{---} \bullet \\ \text{---} \bullet \end{array} \quad (141)$$

for every morphism f .

9.2 Existence of conditional probabilities

Lemma 9.2 (Conditional pushforward). *Suppose we have a sample space (Ω, \mathcal{F}) , variables $\mathsf{X} : \Omega \rightarrow X$ and $\mathsf{Y} : \Omega \rightarrow Y$, $\mathsf{Z} : \Omega \rightarrow Z$ and a probability set $\mathbb{P}_{\{\}}^{\mathsf{X}|\mathsf{Y}}$ with conditional $\mathbb{P}_{\{\}}^{\mathsf{X}|\mathsf{Y}}$ such that $\mathsf{Z} = f \circ \mathsf{Y}$ for some $f : Y \rightarrow Z$. Then there exists a conditional probability $\mathbb{P}_{\{\}}^{\mathsf{Z}|\mathsf{X}} = \mathbb{P}_{\{\}}^{\mathsf{Y}|\mathsf{X}} \mathbb{F}_f$.*

Proof. Note that $(\mathsf{X}, \mathsf{Z}) = (\text{id}_X \otimes f) \circ (\mathsf{X}, \mathsf{Y})$. Thus, by Lemma 4.11, for any $\mathbb{P}_{\alpha} \in \mathbb{P}_{\{\}}$

$$\mathbb{P}_{\alpha}^{\mathsf{XZ}} = \mathbb{P}_{\alpha}^{\mathsf{XY}} \mathbb{F}_{\text{id}_X \otimes f} \quad (142)$$

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

$$\mathbb{F}_{\text{id}_X \otimes f}(A \times B|x, y) = \delta_x(A) \delta_{f(y)}(B) \quad (143)$$

$$= \mathbb{F}_{\text{id}_X}(A|x) \otimes \mathbb{F}_f(B|y) \quad (144)$$

$$\implies \mathbb{F}_{\text{id}_X \otimes f} = \mathbb{F}_{\text{id}_X} \otimes \mathbb{F}_f \quad (145)$$

Thus

$$\mathbb{P}_\alpha^{\mathbf{XZ}} = (\mathbb{P}_\alpha^{\mathbf{X}} \odot \mathbb{P}_{\{\}}^{\mathbf{Y|X}}) \mathbb{F}_{\text{id}_X} \otimes \mathbb{F}_f \quad (146)$$

$$= \begin{array}{c} \text{X} \\ \swarrow \\ \triangleleft \mathbb{P}_\alpha^{\mathbf{X}} \quad \bullet \quad \square \mathbb{P}_{\{\}}^{\mathbf{Y|X}} \quad \square \mathbb{F}_f \quad \longrightarrow \quad \text{Z} \end{array} \quad (147)$$

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 9.3 (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 9.4 (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

$$n(f^{-1}(y)|x, y) = 1 \quad (148)$$

$$m(\mathbf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x, y) m \mathbb{F}_f(dy|x) \forall A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z} \quad (149)$$

In particular, $\mathbb{P}_{\{\}}^{\mathbf{YZ|X}}$ is a Borel measurable map $X \rightarrow Y \times Z$. Thus equation 149 implies for all $A \in \mathcal{Y}$, $B \in \mathcal{Y} \times \mathcal{Z}$

$$\mathbb{P}_{\{\}}^{\mathbf{YZ|X}}(\mathbf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x, y) \mathbb{P}_{\{\}}^{\mathbf{YZ|X}} \mathbb{F}_f(dy|x) \quad (150)$$

$$= \int_A n(B|x, y) \mathbb{P}_{\{\}}^{\mathbf{Y|X}}(dy|x) \quad (151)$$

Where Equation 151 follows from Lemma 9.2.

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

$$\mathbb{P}_\emptyset^{YZ|X}(Y^{-1}(A) \cap B|x) = \int_A n(B|x, y) \mathbb{P}_\alpha^{Y|X}(dy|x) \quad (152)$$

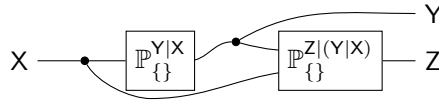
which implies n is a version of $\mathbb{P}_\emptyset^{YZ|(\mathcal{Y}|X)}$. By Lemma 9.2, $n\mathbb{F}_f$ is a version of $\mathbb{P}_\emptyset^{Z|(\mathcal{Y}|X)}$. \square

We might be motivated to ask whether the higher order conditionals in Theorem 9.4 can be chosen to be valid. Despite Lemma 9.9 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 148 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 9.5 (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}_\emptyset with conditional $\mathbb{P}_\emptyset^{YZ|X}$. Then $\mathbb{P}_\emptyset^{Z|(\mathcal{Y}|X)}$ is a version of $\mathbb{P}_\emptyset^{Z|YX}$*

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



$$\mathbb{P}_\alpha^{YZ|X} = \quad (153)$$

$$\Rightarrow \mathbb{P}_\alpha^{XYZ} = \triangleleft \mathbb{P}_\alpha^X \quad \text{---} \quad \text{---} \quad \text{---} \quad \begin{matrix} X \\ Y \\ Z \end{matrix} \quad (154)$$

$$= \triangleleft \mathbb{P}_\alpha^X \quad \text{---} \quad \text{---} \quad \text{---} \quad \begin{matrix} X \\ Y \\ Z \end{matrix} \quad (155)$$

$$= \triangleleft \mathbb{P}_\alpha^{XY} \quad \text{---} \quad \text{---} \quad \text{---} \quad \begin{matrix} X \\ Y \\ Z \end{matrix} \quad (156)$$

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

Theorem 9.6. *Given probability gap model $\mathbb{P}_{\{\}}^{\mathbf{X}, \mathbf{Y}, \mathbf{Z}}$ such that $\mathbb{P}_{\{\}}^{\mathbf{Z}|\mathbf{Y}\mathbf{X}}$ exists, $\mathbb{P}_{\{\}}^{\mathbf{Z}|\mathbf{Y}}$ exists iff $\mathbf{Z} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathbf{X}|\mathbf{Y}$.*

Proof. If: If $\mathbf{Z} \perp\!\!\!\perp_{\mathbb{P}_{\{\}}} \mathbf{X}|\mathbf{Y}$ then by Theorem 6.12, for each $\mathbb{P}_{\alpha} \in \mathbb{P}_{\{\}}$ there exists $\mathbb{P}_{\alpha}^{\mathbf{Z}|\mathbf{Y}}$ such that

$$\mathbb{P}_{\alpha}^{\mathbf{Y}|\mathbf{W}\mathbf{X}} = \begin{array}{c} \mathbf{W} \text{ --- } \boxed{\mathbb{K}} \text{ --- } \mathbf{Y} \\ \mathbf{X} \text{ --- } * \end{array} \quad (157)$$

□

Theorem 9.7 (Valid 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 regular conditional $\mathbb{P}_{\{\}}^{\mathbf{Y}\mathbf{Z}|\mathbf{X}}$, Y discrete and Z standard measurable. Then there exists a valid regular $\mathbb{P}_{\{\}}^{\mathbf{Z}|\mathbf{X}\mathbf{Y}}$.*

Proof. By Theorem 9.4, we have a higher order conditional $\mathbb{P}_{\{\}}^{\mathbf{Z}|\mathbf{Y}(\mathbf{X})}$ which, by Theorem 9.5 is also a version of $\mathbb{P}_{\{\}}^{\mathbf{Z}|\mathbf{X}\mathbf{Y}}$.

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

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

$$\forall A \in \mathcal{Z}, x \in X, y \in Y : \mathbb{P}_{\{\}}^{\mathbf{Y}\mathbf{Z}|\mathbf{X}}(\{y\} \times A|x) = 0 \implies (\mathbb{Q}(A|x, y) = 0) \vee ((\mathbf{X}, \mathbf{Y}) \bowtie \{(x, y)\} = \emptyset) \quad (158)$$

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

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

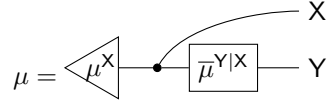
9.3 Validity

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

Definition 9.8 (Regular proper conditional probability). Given a probability space $(\mu, \Omega, \mathcal{F})$ and a variable $\mathbf{X} : \Omega \rightarrow X$, a regular proper conditional probability $\mu^{|\mathbf{X}} : X \rightarrow \Omega$ is Markov kernel such that

$$\mu(A \cap \mathbf{X}^{-1}(B)) = \int_B \mu^{|\mathbf{X}}(A|x) \mu^{\mathbf{X}}(dx) \quad \forall A \in \mathcal{X}, B \in \mathcal{F} \quad (159)$$

$$\iff \quad (160)$$



$$\mu = \triangleleft \mu^{\mathbf{X}} \quad \text{---} \square \bar{\mu}^{\mathbf{Y}|\mathbf{X}} \quad \text{---} \mathbf{Y} \quad \text{---} \mathbf{X} \quad (161)$$

and

$$\mu^{|\mathbf{X}}(\mathbf{X}^{-1}(A)|x) = \delta_x(A) \quad (162)$$

Lemma 9.9. *Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $\mathbf{X} : \Omega \rightarrow X$, $\mathbf{Y} : \Omega \rightarrow Y$, if there is a regular proper conditional probability $\mu^{|\mathbf{X}} : X \rightarrow \Omega$ then there is a valid conditional distribution $\mu^{\mathbf{Y}|\mathbf{X}}$.*

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

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 9.2, \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

$$\mathbb{K}(\mathbf{Y} \bowtie A|x) = \delta_x(\mathbf{Y}^{-1}(A)) \quad (163)$$

$$= 0 \quad (164)$$

Thus \mathbb{K} is valid. \square

Theorem 9.10 (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

$$\mu^{\mathbf{X}}(x) = \mu(\mathbf{X} \bowtie x) \quad (165)$$

$$= \sum_{x' \in X} \mathbb{J}(x') \mathbb{K}(\mathbf{X} \bowtie x|x') \quad (166)$$

$$= 0 \quad (167)$$

$$\neq \mathbb{J}(x) \quad (168)$$

\square

Lemma 9.11 (Semidirect product defines an intersection of probability sets). *Given (Ω, \mathcal{F}) , $X : \Omega \rightarrow (X, \mathcal{X})$, $Y : \Omega \rightarrow (Y, \mathcal{Y})$, $Z : \Omega \rightarrow (Z, \mathcal{Z})$ all standard measurable and maximal probability sets $\mathbb{P}_{\{\}}^{Y|X[M]}$ and $\mathbb{Q}_{\{\}}^{Z|YX[M]}$ then defining*

$$\mathbb{R}_{\{\}}^{YZ|X} := \mathbb{P}_{\{\}}^{Y|X} \odot \mathbb{Q}_{\{\}}^{Z|YX} \quad (169)$$

we have

$$\mathbb{R}_{\{\}}^{YZ|X[M]} = \mathbb{P}_{\{\}}^{Y|X[M]} \cap \mathbb{Q}_{\{\}}^{Z|YX[M]} \quad (170)$$

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

$$\mathbb{R}_a^{XYZ} = \mathbb{R}_a^X \odot \mathbb{P}_{\{\}}^{Y|X} \odot \mathbb{Q}_{\{\}}^{Z|YX} \quad (171)$$

$$\implies \mathbb{R}_a^{XY} = \mathbb{R}_a^X \odot \mathbb{P}_{\{\}}^{Y|X} \quad (172)$$

$$\wedge \mathbb{R}_a^{XYZ} = \mathbb{R}_a^{XY} \odot \mathbb{Q}_{\{\}}^{Z|YX} \quad (173)$$

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

Lemma 9.12 (Equivalence of validity definitions). *Given $X : \Omega \rightarrow X$, with Ω and X standard measurable, a probability measure $\mathbb{P}^X \in \Delta(X)$ is valid if and only if the conditional $\mathbb{P}^{X|*} := * \mapsto \mathbb{P}^X$ is valid.*

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

$$\forall A \in \mathcal{X} : X \bowtie A = \emptyset \implies \mathbb{P}^{X|*}(A|*) = 0 \quad (174)$$

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

$$\forall A \in \mathcal{X} : X \bowtie A = \emptyset \implies \mathbb{P}^X(A) = 0 \quad (175)$$

\square

Lemma 9.13 (Semidirect product of valid candidate conditionals is valid). *Given (Ω, \mathcal{F}) , $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, $Z : \Omega \rightarrow Z$ (all spaces standard measurable) and any valid candidate conditional $\mathbb{P}^{Y|X}$ and $\mathbb{Q}^{Z|YX}$, $\mathbb{P}^{Y|X} \odot \mathbb{Q}^{Z|YX}$ is also a valid candidate conditional.*

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

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

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

$$\mathbb{R}^{YZ|X}(B \times C|x) = \int_B \mathbb{Q}^{Z|YX}(C|y, x) \mathbb{P}^{Y|X}(dy|x) \quad (176)$$

$$\leq \mathbb{P}^{Y|X}(B|x) \quad (177)$$

$$= 0 \quad (178)$$

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

$$\mathbb{R}^{YZ|X}(B \times C|x) = \int_B \mathbb{Q}^{Z|YX}(C|y, x) \mathbb{P}^{Y|X}(dy|x) \quad (179)$$

$$= 0 \quad (180)$$

□

Corollary 9.14 (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 9.13 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 9.12. □

Theorem 9.15 (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 9.14.

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

$$(\mathbb{V}^X \odot \mathbb{T}^{Y|X})(x, y) = \mathbb{T}^{Y|X}(y|x) \mathbb{V}^X(x) \quad (181)$$

$$= \mathbb{T}^{Y|X}(y|x) \quad (182)$$

$$> 0 \quad (183)$$

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

9.4 Conditional independence

Theorem 6.13. *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*

$$\mathbb{P}_C^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (184)$$

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

$$\mathbb{P}_\beta^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (185)$$

And so, by Theorem 6.12, $Y \perp\!\!\!\perp_{\mathbb{P}_\beta} X|W$ for all $\beta \in A$, and in particular $Y \perp\!\!\!\perp_{\mathbb{P}_\alpha} X|W$. Only if: By Theorem 6.12, there exists a version of $\mathbb{P}_\alpha^{Y|WX}$ such that

$$\mathbb{P}_\alpha^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{P}_\alpha^{Y|W}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (186)$$

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

$$\mathbb{P}_C^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{P}_\alpha^{Y|W}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (187)$$

□

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 $Y \perp\!\!\!\perp_{\mathbb{P}_\square} X$ 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 $\mathbb{P}_\square^{Y|WX}$, $Y \perp\!\!\!\perp_{\mathbb{P}_\square} X|W$ 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 $Y \perp\!\!\!\perp_{\mathbb{P}} X$ but \mathbb{P}^Y does not exist.

9.5 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 9.16 (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 9.17 (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:

$$\forall B \in \mathcal{Y}, x \in X : (X, Y) \bowtie \{x\} \times B = \emptyset \implies (\mathbb{L}(B|x) = 0) \vee (X \bowtie \{x\} = \emptyset) \quad (188)$$

Definition 9.18 (Maximal probability set). Given (Ω, \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 9.13 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 9.15).

Example 9.19. 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 $\mathcal{B} = \frac{W}{H^2}$ - 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

$$\mathbb{P}_{\square}^{B|WHC} = \begin{array}{c} H \text{ ---} * \\ C \text{ ---} \text{-----} B \\ W \text{ ---} * \end{array} \quad (189)$$

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 $\mathbb{P}_{\square}^{B|WHC}(\{x\}|w, h, x) = 1$, but

$$(B, W, H) \bowtie \{(x, w, h)\} = \{\omega | (W, H)(\omega) = (w, h), B(\omega) = \frac{w}{h^2}\} \quad (190)$$

$$= \emptyset \quad (191)$$

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

Validity rules out conditional probabilities like 189. 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 189 would be used to evaluate the causal effect of body mass index in the causal diagram found in Shahar (2009).

9.6 Causal contractibility

Theorem 7.4. *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

$$\mathbb{P}_C^{Y|D}(y_1, y_2 | d_1, d_2) = \llbracket (y_1, y_2) = (d_1 + d_2, d_1 + d_2) \rrbracket \quad (192)$$

then

$$\mathbb{P}_C^{Y_1|D}(y_1 | d_1, d_2) = \llbracket y_1 = d_1 + d_2 \rrbracket \quad (193)$$

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 $\{0, 1\} \rightarrow \{0, 1\}$

$$\text{swap}_{\rho(D)} \mathbb{P}_C^{Y|D}(y_1, y_2 | d_1, d_2) = \mathbb{P}_C^{Y|D}(y_1, y_2 | d_2, d_1) \quad (194)$$

$$= \llbracket (y_1, y_2) = (d_2 + d_1, d_2 + d_1) \rrbracket \quad (195)$$

$$= \llbracket (y_1, y_2) = (d_1 + d_2, d_1 + d_2) \rrbracket \quad (196)$$

$$= \llbracket (y_2, y_1) = (d_1 + d_2, d_1 + d_2) \rrbracket \quad (197)$$

$$= \mathbb{P}_C^{Y|D} \text{swap}_{\rho(Y)}(y_1, y_2 | d_1, d_2) \quad (198)$$

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

$$\mathbb{P}_C^{Y|D}(y_1, y_2 | d_1, d_2) = \llbracket (y_1, y_2) = (0, 1) \rrbracket \quad (199)$$

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

$$\mathbb{P}_C^{\mathbf{Y}_1|\mathbf{D}}(y_1|d_1, d_2) = \mathbb{I}[y_1 = 0] \quad (200)$$

$$= \mathbb{P}_C^{\mathbf{Y}_1|\mathbf{D}_1}(y_1|d_1) \quad (201)$$

$$\mathbb{P}_C^{\mathbf{Y}_2|\mathbf{D}}(y_2|d_1, d_2) = \mathbb{I}[y_2 = 1] \quad (202)$$

$$= \mathbb{P}_C^{Y_2|D_2}(y_2|d_2) \quad (203)$$

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

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

$$\text{swap}_{\rho(D)} \mathbb{P}_C^{\mathbf{Y}|D}(y_1, y_2 | d_1, d_2) = \mathbb{P}_C^{\mathbf{Y}|D}(y_1, y_2 | d_2, d_1) \quad (204)$$

$$= \mathbb{I}(y_1, y_2) = (0, 1) \mathbb{I} \quad (205)$$

$$\neq \mathbb{I}(y_2, y_1) = (0, 1) \mathbb{I} \quad (206)$$

$$= \mathbb{P}_C^{Y|D} \text{swap}_{\rho(D)}(y_1, y_2 | d_1, d_2) \quad (207)$$

1

Theorem 7.7. *Given a probability set \mathbb{P}_C such that $D := (D_i)_{i \in \mathbb{N}}$ and $Y := (Y_i)_{i \in \mathbb{N}}$, \mathbb{P}_C is $(D; Y)$ -causally contractible if and only if there exists a column exchangeable probability distribution $\mu^{Y^D} \in \Delta(Y^{D \times \mathbb{N}})$ such that*

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \triangle \\ \mu^{Y^D} \\ D - [F_{ev}] - Y \end{array} \quad (208)$$

$$\Longleftrightarrow \quad (209)$$

$$\mathbb{P}_C^{\mathbf{Y}|\mathbf{D}}(y|(d_i)_{i \in \mathbb{N}}) = \mu^{\mathbf{Y}^D} \Pi_{(d_i)_{i \in \mathbb{N}}}(y) \quad (210)$$

Where $\Pi_{(d_i, i)_{i \in \mathbb{N}}} : Y^{|D| \times \mathbb{N}} \rightarrow Y^{\mathbb{N}}$ 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

$$ev: D^{\mathbb{N}} \times Y^{D \times \mathbb{N}} \rightarrow Y \quad (211)$$

$$((d_i)_{\mathbb{N}}, (y_{ij})_{i \in D, j \in \mathbb{N}}) \mapsto (y_{di})_{i \in \mathbb{N}} \quad (212)$$

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 $\mathbb{P}_{C'}^{\mathbf{Y}|D} \stackrel{P_C}{\cong} \mathbb{P}_C^{\mathbf{Y}|D}$. It can be constructed by adding to \mathbb{P}_C probability sets with marginals $\delta_d \odot \mathbb{P}_{C'}^{\mathbf{Y}|D}$ 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

$$\mathbb{P}_\alpha^{\text{DY}}(B \times C) = \int_B \mathbb{P}_C^{\text{Y|D}}(C|d) \mathbb{P}_\alpha^{\text{D}}(dd) \quad (213)$$

$$= \int_B \int_D \mathbb{P}_C^{\mathbf{Y}|D}(C|d') \mathbb{P}_d^D(dd') \mathbb{P}_\alpha^D(dd) \quad (214)$$

$$= \int_B \mathbb{P}_d^Y(C) \mathbb{P}_\alpha^D(dd) \quad (215)$$

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

$$\mu^{Y^D}((y_{ij})_{D \times \mathbb{N}}) := \mathbb{P}_e^Y((y_{|D|i+j})_{i \in D, j \in \mathbb{N}}) \quad (216)$$

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

Define

$$\mathbb{Q} : D \rightarrowtail Y \quad (217)$$

$$Q := \begin{array}{c} \triangle \\ \mu^{Y^D} \\ D \text{ --- } \boxed{\text{F}_{\text{ev}}} \text{ --- } Y \end{array} \quad (218)$$

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

$$\sum_{y \in Y^{-1}(y_A)} \mathbb{Q}(y|d) = \sum_{y \in Y^{-1}(y_A)} \mu^{(Y_{d_{ii}}^D)^A}(y) \quad (219)$$

$$= \sum_{y \in Y^{-1}(y_A)} \mathbb{P}_e^{(Y|D|d_i+i)A}(y) \quad (220)$$

$$= \mathbb{P}_e^{\mathbf{Y}_B}(y_A) \quad (221)$$

$$= \mathbb{P}_d^{\mathbf{Y}^A}(y_A) \quad \text{by causal contractibility} \quad (222)$$

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

$$\mathbb{Q}(y|d) = \mathbb{P}_d^{\mathbf{Y}}(y) \quad (223)$$

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

$$\mu^{Y^D} \Pi_S = \mathbb{P}_e^{Y_{\rho(S)}} \quad (224)$$

$$= \mathbb{P}_e^{Y_{\rho(T)}} \quad \text{by contractibility of } \mathbb{P}_C \text{ and the equality } e_{\rho(S)} = e_{\rho(T)} \quad (225)$$

$$= \mu^{Y^D} \Pi_T \quad (226)$$

If: Suppose

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \triangle \\ \mu^{Y^D} \\ \text{D} \end{array} \text{---} \boxed{\mathbb{F}_{\text{ev}}} \text{---} Y \quad (227)$$

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

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

By definition,

$$\mathbb{P}_C^{Y_S|D}(y_S|d) = \sum_{y_S^D \in Y^{D|D| \times |S|}} \mu^{Y^D} \Pi_S(y_S^D) \mathbb{F}_{\text{ev}}(y_S|d, y_S^D) \quad (228)$$

$$= \sum_{y_S^D \in Y^{D|D| \times |T|}} \mathbb{P}_C^{Y_T^D}(y_S^D) \mathbb{F}_{\text{ev}}(y_S|d, y_S^D) \quad \text{by contractibility of } \mu^{Y^D} \Pi_T \quad (229)$$

$$= \mathbb{P}_C^{Y_T|D}(y_S|d) \quad (230)$$

□

Theorem 7.9. 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

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \triangle \\ \mu^H \\ \text{D} \end{array} \text{---} \boxed{\begin{array}{c} \Pi_{D,i} \text{---} \boxed{\mathbb{P}_{\square}^{Y_0|HD_0}} \text{---} Y_i \\ i \in \mathbb{N} \end{array}} \quad (231)$$

$$\iff \quad (232)$$

$$Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e Y_{\mathbb{N} \setminus i}, D_{\mathbb{N} \setminus i} C | HD_i \quad \forall i \in \mathbb{N} \quad (233)$$

$$\wedge H \perp\!\!\!\perp_{\mathbb{P}_C}^e DC \quad (234)$$

$$\wedge \mathbb{P}_C^{Y_i|HD_i} = \mathbb{P}^{Y_0|HD_0} \quad \forall i \in \mathbb{N} \quad (235)$$

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

Proof. We make use of Lemma 7.7 to show that we can represent the conditional probability $\mathbb{P}_C^{Y|D}$ as

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \triangle \mu^{Y^D} \\ \text{D} \longrightarrow \text{F}_{\text{ev}} \longrightarrow Y \end{array} \quad (236)$$

$$(237)$$

As a preliminary, we will show

$$\mathbb{F}_{\text{ev}} = \begin{array}{c} \text{H} \longrightarrow \text{D} \longrightarrow \begin{array}{c} \Pi_{Y^D, i} \longrightarrow \text{F}_{\text{ev}, i} \longrightarrow Y_i \\ \Pi_{D, i} \longrightarrow \text{F}_{\text{ev}, i} \longrightarrow Y_i \end{array} \end{array} \quad (238)$$

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

$$((y_i)_{i \in D}, d) \mapsto y_d \quad (239)$$

Recall that ev is the function

$$((d_i)_{i \in \mathbb{N}}, (y_{ij})_{i \in D, j \in \mathbb{N}}) \mapsto (y_{d_i i})_{i \in \mathbb{N}} \quad (240)$$

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

$$\mathbb{F}_{\text{ev}}\left(\prod_{i \in \mathbb{N}} A_i | (d_i)_{i \in \mathbb{N}}, (y_{ij})_{i \in D, j \in \mathbb{N}}\right) = \delta_{(y_{d_i i})_{i \in \mathbb{N}}} \left(\prod_{i \in \mathbb{N}} A_i\right) \quad (241)$$

$$= \prod_{i \in \mathbb{N}} \delta_{y_{d_i i}}(A_i) \quad (242)$$

$$= \text{copy}^{\mathbb{N}} \prod_{i \in \mathbb{N}} (\Pi_{D, i} \otimes \Pi_{Y, i}) \mathbb{F}_{\text{ev}_{Y^D \times D}} \quad (243)$$

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 7.7). From kal (2005) we have a directing random measure H such that

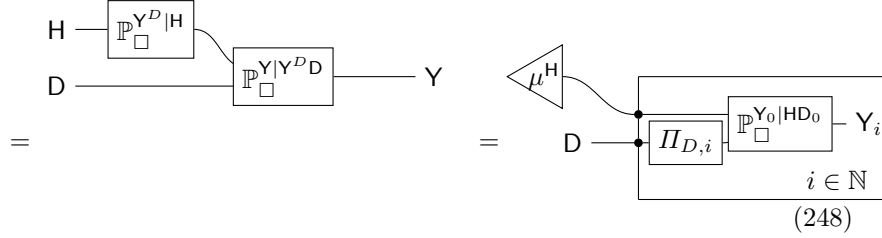
$$\mathbb{P}_C^{Y^D | H} = \begin{array}{c} \text{H} \longrightarrow \text{D} \longrightarrow \begin{array}{c} \mathbb{P}_C^{Y^D | H} \longrightarrow Y_i \end{array} \end{array} \quad (244)$$

$$\iff \quad (245)$$

$$\mathbb{P}_C^{Y^D | H} \left(\prod_{i \in \mathbb{N}} A_i | h \right) = \prod_{i \in \mathbb{N}} \mathbb{P}_C^{Y_i^D | H} (A_i | h) \quad (246)$$

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

$$\mathbb{P}_C^{Y|HD} = \mathbb{P}_C^{Y^D|HD} \odot \mathbb{P}_C^{Y|Y^DHD} \quad (247)$$



If: By assumption

$$\mathbb{P}_C^{Y|D}(\prod_{i \in \mathbb{N}} A_i | h, (d_i)_{i \in \mathbb{N}}) = \int_H \prod_{i \in \mathbb{N}} \mathbb{P}_C^{Y_i|HD_1}(A_i | h, d_i) \mathbb{P}_C^H(dh) \quad (249)$$

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

$$\mathbb{P}_\alpha^{Y_M}(A) = \int_{D^\mathbb{N}} \mathbb{P}_\alpha^{Y_M|D}(A|d) \mathbb{P}_\alpha^D(dd) \quad (250)$$

$$= \int_H \int_{D^\mathbb{N}} \prod_{i \in M} \mathbb{P}_C^{Y_i|HD_1}(A_i | h, d_i) \mathbb{P}_\alpha^D(dd) \mathbb{P}_C^H(dh) \quad (251)$$

$$= \int_H \int_{D^{|M|}} \prod_{i \in M} \mathbb{P}_C^{Y_i|HD_1}(A_i | h, d_i) \mathbb{P}_\alpha^{D_M}(dd_M) \mathbb{P}_C^H(dh) \quad (252)$$

$$= \int_H \int_{D^{|M|}} \prod_{i \in M} \mathbb{P}_C^{Y_i|HD_1}(A_i | h, d_i) \mathbb{P}_{\alpha'}^{D_N}(dd_N) \mathbb{P}_C^H(dh) \quad (253)$$

$$= \int_H \int_{D^\mathbb{N}} \prod_{i \in M} \mathbb{P}_C^{Y_i|HD_1}(A_i | h, d_i) \mathbb{P}_{\alpha'}^D(dd) \mathbb{P}_C^H(dh) \quad (254)$$

$$= \mathbb{P}_{\alpha'}^{Y_M}(A) \quad (255)$$

□

9.7 Body mass index revisited

Lemma 9.20. *Suppose we have a probability set \mathbb{P}_C that is $(D, X; Y)$ -causally contractible where $D := (D_i)_{i \in M}$ and similarly for X and Y . If $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e D_i C | X_i H$, then \mathbb{P}_C is also $(X; Y)$ -causally contractible.*

Proof. From causal contractibility we have

$$Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{\{i\}^c}, X_{\{i\}^c}, D_{\{i\}^c}) C | HD_i X_i \quad (256)$$

Combining this with the assumption $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e D_i C | X_i H$ we have, by contraction,

$$Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{\{i\}^c}, X_{\{i\}^c}) C | H X_i \quad (257)$$

Furthermore, also from causal contractibility, for all $i, j \in M$

$$\mathbb{P}_C^{Y_i | X_i D_i H} \cong \mathbb{P}_C^{Y_j | X_j D_j H} \quad (258)$$

$$\implies \mathbb{P}_C^{Y_i | X_i H} \cong \mathbb{P}_C^{Y_j | X_j H} \quad (259)$$

□

Theorem 7.11. *Suppose we have a probability set \mathbb{P}_C that is $(D; X, Y)$ -causally contractible, where $D := (D_i)_{i \in M}$ and similarly for X and Y . If there exists $\alpha \in C$ such that $\mathbb{P}_\alpha^D \gg \{\mathbb{P}_\beta^D | \beta \in C\}$ 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. By Corollary 6.14 and the existence of $\mathbb{P}_C^{Y_i X_i | H D_i}$ for all $i \in M$, $\mathbb{P}_C^{Y_i | H X_i D_i}$ also exists for all i . Furthermore, because $\mathbb{P}_C^{Y_i X_i | H D_i} = \mathbb{P}_C^{Y_j X_j | H D_j}$ for all $i, j \in M$, $\mathbb{P}_C^{Y_i | H X_i D_i} = \mathbb{P}_C^{Y_j | H X_j D_j}$ for all $i, j \in M$.

From causal contractibility we have

$$(X_i, Y_i) \perp\!\!\!\perp_{\mathbb{P}_\alpha} (X_{\{i\}^c}, Y_{\{i\}^c}, D_{\{i\}^c}) | H D_i \quad (260)$$

$$Y_i \perp\!\!\!\perp_{\mathbb{P}_\alpha} (Y_{\{i\}^c}, X_{\{i\}^c}) | H D_i X_i \quad (261)$$

Where Eq. 261 follows from 260 by weak union. By Theorem 6.13, $Y_i \perp\!\!\!\perp_{\mathbb{P}_\alpha} (Y_{\{i\}^c}, X_{\{i\}^c}) | H D_i X_i$ also, and so \mathbb{P}_C is $(D, X; Y)$ -causally contractible.

$Y_i \perp\!\!\!\perp_{\mathbb{P}_\alpha} D_i | (H, X_i)$ for all $i \in M$ implies $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e D_i C | (H, X_i)$ for all $i \in M$ by Theorem 6.13. The result follows by noting that \mathbb{P}_C is also $(D, X; Y)$ -causally contractible by higher order conditionals, and therefore $(X; Y)$ -causally contractible by Lemma 9.20. □

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 Dufo 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 Dufo, 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.
- Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, January 2017. ISSN 1349-6964. doi: 10.1007/s41237-016-0008-2. URL <https://doi.org/10.1007/s41237-016-0008-2>.
- 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.
- Frederick Eberhardt. A contemporary example of Reichenbachian coordination. *Synthese*, 200(2):90, March 2022. ISSN 1573-0964. doi: 10.1007/s11229-022-03571-8. URL <https://doi.org/10.1007/s11229-022-03571-8>.
- 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>.
- Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Electronic edition, 3rd edition edition, 2021. URL <http://www.stat.columbia.edu/~gelman/book/BDA3.pdf>.
- 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>.
- Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, 2015. ISBN 978-0-521-88588-1. doi: 10.1017/CBO9781139025751. URL <https://www.cambridge.org/core/books/causal-inference-for-statistics-social-and-biomedical-sciences/71126BE90C58F1A431FE9B2DD07938AB>.
- 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.
- 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.
- Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, June 2020. ISSN 0090-5364, 2168-8966. doi: 10.1214/19-AOS1857. URL <https://projecteuclid.org/journals/annals-of-statistics/volume-48/issue-3/The-hardness-of-conditional-independence-testing-and-the-generalised-covariance/10.1214/19-AOS1857.full>. Publisher: Institute of Mathematical Statistics.
- Eyal Shahrar. The association of body mass index with health outcomes: causal, inconsistent, or confounded? *American Journal of Epidemiology*, 170(8): 957–958, October 2009. ISSN 1476-6256. doi: 10.1093/aje/kwp292.
- Peter Spirtes and Richard Scheines. Causal Inference of Ambiguous Manipulations. *Philosophy of Science*, 71(5):833–845, December 2004. ISSN 0031-8248, 1539-767X. doi: 10.1086/425058. URL <https://www.cambridge.org/core/journals/philosophy-of-science/article/abs/causal-inference-of-ambiguous-manipulations/2A605BCFFC1A879A157966473AC2A6D2>. Publisher: Cambridge University Press.
- 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=tv&paperid=4710&option_lang=eng.
- Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.
- James Woodward. Causation and Manipulability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016. URL <https://plato.stanford.edu/archives/win2016/entries/causation-mani/>.

Appendix: