

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

David Johnston, Cheng Soon Ong, Robert C. Williamson

March 29, 2022

Contents

1	Introduction	2
1.1	Our approach	3
1.2	Paper Outline	4
2	Decision problems	5
2.1	Other decision theoretic causal models	6
3	Probability prerequisites	7
3.1	Variables	7
3.2	Other decision theoretic causal models	7
3.3	Standard probability theory	8
3.4	String diagram notation	10
3.5	Products	10
3.6	Elements of string diagrams	11
3.7	Iterated copy maps and plates	12
3.7.1	Examples	13
3.8	Probability sets	14
3.9	Extended conditional independence	16
3.10	Examples	17
4	When do response functions exist?	19
4.1	Relevance to previous work	20
4.2	Causal contractibility	21
4.3	Existence of response conditionals	23
4.4	Elaborations and examples	25
4.5	Assessing causal contractibility	26
4.6	Body mass index revisited	30
5	Conclusion	31
5.1	Choices aren't always known	32

6	Appendix, needs to be organised	33
6.1	Markov categories	33
6.2	Existence of conditional probabilities	34
6.3	Validity	37
6.4	Conditional independence	40
6.5	Maximal probability sets and valid conditionals	41
6.6	Causal contractibility	43
6.7	Body mass index revisited	50

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, and as a result, they leave us searching in vain for “elementary manipulations”. 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 can prove two basic but important facts about when probabilistic causal effects exist.

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 map* (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 not every pair of variables admits a unique interventional probability map or unique potential outcomes. In the potential outcomes framework, the “stable unit-treatment value assumption” (SUTVA) has been offered as a sufficient condition for the existence of potential outcomes (Rubin (2005); Imbens and Rubin (2015)). 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

We tackle the problem of when “causal effects” exist. Our approach differs from one or both of the popular approaches discussed above in a number of ways:

1. We assume that the objective analysis is to compare different choices available in a decision problem
2. Variables are associated with measurement procedures, not generic “kinds” of measurement procedures

3. Probability distributions represent uncertain knowledge about the results of measurement procedures

We will explain these three features in some more detail. First, our approach is decision theoretic: as Dawid (2021) and Heckerman and Shachter (1995) have observed, when we are faced with a decision problem there is automatically a presumptive set of “elementary interventions” – namely, the different choices that we want to compare. This set is exactly the right size: we need to compare every choice available, and there is no comparison that we need to make involving anything but the choices available. This assumption sets our approach apart from variations of potential outcomes and causal graphical models in which causal relationships are tied to random variables.

Second, variables in our approach are associated with concrete measurement procedures. Interventional causal models typically model relationships between what Dawid (2021) calls “generic variables”. Generic variables represent “types” of things we can measure, not the results of particular measurements. In standard statistical analysis, generic variables “ X ” and “ Y ” might be used to refer to a model of an independent and identically distributed sequence of variables $(X_1, Y_1), (X_2, Y_2), \dots$. As we are interested in when such “generic causal relationships” exist, we take as a starting point variables that represent the results of particular measurements, not the results of “types” of measurement.

Thirdly, we regard probabilistic causal models as models of our uncertain knowledge of the outcomes of the different choices facing us. This differs from some approaches to causal graphical models that look for causal relationships “in nature”. In Section 4.5, for example, we discuss how causal conclusions can be motivated by a judgement that, no matter which choice we make, our resulting model is indifferent to the ordering of certain “index variables”.

These differences do not mean that our approach enables any formal results that cannot be found in any other approach. In fact, we speculate the opposite is true: probability functions underpin our approach, causal Bayesian networks, structural causal models, single world intervention graphs or any of various approaches to decision theoretic causal modelling are all probability functions, and there are a number of reasonable ways to extend potential outcomes models to probability functions. Thus, anything that can be said with probability functions in our approach can likely be said using probability functions in any of these other approaches. However, because we have a slightly different understanding of what these probability functions represent, we end up finding different results.

1.2 Paper Outline

The key results of this paper are in Section 4. In particular, Theorem 4.9 establishes *causal contractibility* as a necessary and sufficient condition for the existence of *response functions*. Response functions are the name we use for the particular kind of “probabilistic causal effects” investigated in this paper. Theorem 4.13 is relevant to justifying causal contractibility in some cases of ran-

domised experiments and active choice. To our knowledge, existing arguments about the identifiability of causal effects tend to be somewhat informal and do not also pertain to active choices. Theorem 4.13 is notably equally applicable to randomised experiments and active choice, and makes substantially different assumptions to existing results of this nature.

To prove these theorems, we need some standard probability theory, introduced in Section 3. We also need to extend standard probability theory to a theory of *probability sets*, which we introduce in Section 3.8. The theory of probability sets is in many ways similar to standard probability theory, except instead of conditional probability distributions we talk about *uniform conditional probability*, which does not always exist, and instead of conditional independence we use *extended conditional independence* as introduced by Constantinou and Dawid (2017).

We also make use of a graphical language to represent probability sets. Our notation was created for reasoning about abstract Markov categories, and is somewhat different to existing graphical languages. The main difference is that in our notation wires represent variables and boxes (which are like nodes in directed acyclic graphs) represent probabilistic functions. Standard directed acyclic graphs annotate nodes with variable names and represent probabilistic functions implicitly. The advantage of explicitly representing probabilistic functions is that we can write equations involving graphics. This is not critical to understanding our approach, and we offer statements of all definitions and results in more standard notation, though some proofs depend on the graphical notation. This is introduced in Section 3.4.

2 Decision problems

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

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

We make the assumption that each choice is associated with a measure-

ment 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*.

2.1 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 3.24, on the basis of Theorem 3.26.

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 ??), 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*.

3 Probability prerequisites

Notation table, including iverson bracket

3.1 Variables

Our definition of variables is very similar to the standard definition given in many statistics textbooks. Suppose we have a measurement procedure where we interact with the real world somehow, and from this interaction we end up with a vector of numbers. Once we have this vector, we can apply functions to it – project the first element, add the first to the second and so forth. We say that the set of elements that the measurement process can give, along with a σ -algebra of events, (Ω, \mathcal{F}) is the *sample space* associated with the measurement procedure, and each measurable function $f : \Omega \rightarrow Y$ for some measurable (Y, \mathcal{Y}) is a *variable*. The measurement procedure plays the crucial role of translating the real world into numbers that we can operate on mathematically, and a *model* is a probability measure on (Ω, \mathcal{F}) that represents in the world of mathematics our understanding of the real world things addressed by the measurement process.

However, we assume here that we’re actually interested in constructing models to help us make decisions. We consider this means we have a set of available choices C , and with each $\alpha \in C$ we may have a *different* measurement procedure, though we hold that they all yield values in the same sample space (Ω, \mathcal{F}) . With each measurement procedure, we suppose we have a probability measure \mathbb{P}_α on (Ω, \mathcal{F}) , so all together we have a probability function $\alpha \mapsto \mathbb{P}_\alpha$. We are particularly interested in the set of all of the probability measures $\mathbb{P}_C : \{\mathbb{P}_\alpha | \alpha \in C\}$ – anything that is true of every element of this set is true regardless of what we ultimately decide.

With probability sets we can define *uniform conditional distributions* (Definition 3.18) and *extended conditional independence* (Definition 3.24; see also Constantinou and Dawid (2017)) which are similar but not identical to standard notions of conditional distributions and conditional independence.

Section 3.3 introduces standard probability theory, Section 3.4 introduces the string diagram notation that we use to represent Markov kernels (or probabilistic functions) and Section 3.8 introduces additional theory specific to probability sets.

3.2 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 *CausalBayesCon-*

struct 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 3.24, on the basis of Theorem 3.26.

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 ??), 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*.

3.3 Standard probability theory

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

We use a number of shorthands for measurable spaces:

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

Definition 3.2 (Probability measure). Given a measurable space (X, \mathcal{X}) , a probability measure is a σ -additive function $\mu : \mathcal{X} \rightarrow [0, 1]$ 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 3.3 (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 3.4 (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 3.5 (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 3.6 (Variable). Given a measurable space (Ω, \mathcal{F}) and a measurable space of values (X, \mathcal{X}) , an X -valued variable is a measurable function $X : (\Omega, \mathcal{F}) \rightarrow (X, \mathcal{X})$.

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

Definition 3.8 (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 3.9 (Distribution-kernel products). Given (X, \mathcal{X}) , (Y, \mathcal{Y}) a probability distribution $\mu \in \Delta(X)$ and a Markov kernel $\mathbb{K} : X \rightarrow Y$, $\mu\mathbb{K}$ is a probability distribution on (Y, \mathcal{Y}) defined by

$$\mu\mathbb{K}(A) := \int_X \mathbb{K}(A|x)\mu(dx) \quad (1)$$

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

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

$$\mathbb{K}\mathbb{L}(A|x) := \int_Y \mathbb{L}(A|y)\mathbb{K}(dy|x) \quad (2)$$

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

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

$$\mu^X = \mu\mathbb{F}_X \quad (3)$$

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

$$\mu\mathbb{F}_X(A) = \int_\Omega \delta_{X(\omega)}(A)d\mu(\omega) \quad (4)$$

$$= \int_{X^{-1}(A)} d\mu(\omega) \quad (5)$$

$$= \mu^X(A) \quad (6)$$

□

3.4 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).

3.5 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 3.12 (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 $A \in \mathcal{Y}$

$$\mu\mathbb{K}(A) = \int_X \mathbb{K}(A|x)\mu(dx) \quad (7)$$

Definition 3.13 (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 $B \in \mathcal{Z}$

$$\mathbb{K}\mathbb{L}(B|x) = \int_Y \mathbb{L}(B|y)\mathbb{K}(dy|x) \quad (8)$$

We can also define a tensor product of kernels.

Definition 3.14 (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

$$\mathbb{K} \otimes \mathbb{M}(A \times B|x, w) = \mathbb{K}(A|x)\mathbb{M}(B|w) \quad (9)$$

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

3.6 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:

$$\mathbb{K} := \text{---} \boxed{\mathbb{K}} \text{---} \quad (10)$$

$$\mu := \triangleleft \boxed{\mathbb{P}} \text{---} \quad (11)$$

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:

$$\mathbb{L}\mathbb{M} := X \boxed{\mathbb{K}} \text{---} \boxed{\mathbb{M}} Z \quad (12)$$

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:

$$\mathbb{K} \otimes \mathbb{L} := \begin{array}{c} W \boxed{\mathbb{K}} Y \\ X \boxed{\mathbb{L}} Z \end{array} \quad (13)$$

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:

$$\text{id}_X := X \text{---} X \quad (14)$$

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:

$$\text{del}_X := \text{---} * X \quad (15)$$

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:

$$\text{copy}_X := X \text{---} \begin{array}{c} X \\ X \end{array} \quad (16)$$

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:

$$\text{swap}_{X,Y} := \begin{array}{c} \diagup \quad \diagdown \\ \diagdown \quad \diagup \end{array} \quad (17)$$

Diagrams in Markov categories satisfy the commutative comonoid axioms (see Definition 6.1)

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} \nearrow \\ \searrow \end{array} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \\ \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \end{array} = \begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \\ \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \end{array} \quad (18)$$

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \bullet \\ \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \bullet \end{array} \begin{array}{c} \text{---} \\ \text{---} \end{array} = \begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \bullet \\ \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \bullet \end{array} \begin{array}{c} \text{---} \bullet \\ \text{---} \bullet \end{array} \quad (19)$$

$$\begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \\ \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \end{array} = \begin{array}{c} \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \bullet \\ \text{---} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \bullet \end{array} \quad (20)$$

as well as compatibility with the monoidal structure

$$X \otimes Y \xrightarrow{\quad} \bullet = \begin{array}{c} X \xrightarrow{\quad} \bullet \\ X \xrightarrow{\quad} \bullet \end{array} \quad (21)$$

$$X \otimes Y \xrightarrow{\quad} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} X \otimes Y \\ X \otimes Y \end{array} = \begin{array}{c} X \xrightarrow{\quad} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} X \\ Y \end{array} \\ Y \xrightarrow{\quad} \bullet \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} X \\ Y \end{array} \end{array} \quad (22)$$

and the naturality of del , which means that

$$\begin{array}{c} \text{---} \boxed{f} \text{---} \bullet \\ \text{---} \bullet \end{array} = \begin{array}{c} \text{---} \bullet \\ \text{---} \bullet \end{array} \quad (23)$$

3.7 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 3.15 (n -fold copy map). The n -fold copy map $\text{copy}_X^n : X \rightarrow X^n$ is given by

$$\text{copy}_X^1 = \text{copy}_X \quad (24)$$

$$\text{copy}_X^n = \begin{array}{c} \boxed{\text{copy}_X^{n-1}} \\ \text{---} \bullet \text{---} \end{array} \quad n > 1 \quad (25)$$

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 $i \in A$,

$$\bigotimes_{i \in A} \mathbb{K}_i := \boxed{\begin{array}{c} \boxed{\mathbb{K}_i} \\ i \in A \end{array}} \text{copy}_X^{|A|} \left(\bigotimes_{i \in A} \mathbb{K}_i \right) := \text{---} \bullet \boxed{\begin{array}{c} \boxed{\mathbb{K}_i} \\ i \in A \end{array}} \quad (26)$$

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

$$\begin{array}{c} \text{---} \boxed{\mathbb{K}} \text{---} \\ \text{---} \boxed{\mathbb{L}} \text{---} \bullet \end{array} = (\text{id}_X \otimes \mathbb{L})\mathbb{K} \quad (27)$$

$$[(\text{id}_X \otimes \mathbb{L})\mathbb{K}](A|x, w) = \int_Y \int_X \mathbb{K}(A|x', y') \mathbb{L}(dy'|w) \delta_x(dx') \quad (28)$$

$$= \int_Y \mathbb{K}(A|x, y') \mathbb{L}(dy'|w) \quad (29)$$

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

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

$$\begin{array}{c} \text{---} \boxed{\mathbb{K}} \text{---} \\ \text{---} \bullet \text{---} \end{array} = (\text{id}_X \otimes \text{copy}_Y)\mathbb{K} \quad (30)$$

$$[(\text{id}_X \otimes \text{copy}_Y)\mathbb{K}](A|x, y) = \int_Y \int_Y \mathbb{K}(A|x, y', y'') \delta_y(dy') \delta_y(dy'') \quad (31)$$

$$= \mathbb{K}(A|x, y, y) \quad (32)$$

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

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

$$\begin{array}{c} \text{---} \times \text{---} \text{---} \boxed{\mathbb{K}} \text{---} \\ = \text{swap}_{YX} \mathbb{K} \end{array} \quad (33)$$

$$(\text{swap}_{YX} \mathbb{K})(A|y, x) = \int_{X \times Y} \mathbb{K}(A|x', y') \delta_y(dy') \delta_x(dx') \quad (34)$$

$$= \mathbb{K}(A|x, y) \quad (35)$$

The swap map before a kernel switches the input arguments.

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

$$\begin{array}{c} \text{---} \boxed{\mathbb{K}} \times \text{---} \\ = \mathbb{K} \text{swap}_{YZ} \end{array} \quad (36)$$

$$(\mathbb{K} \text{swap}_{YZ})(A \times B|x) = \int_{Y \times Z} \delta_y(B) \delta_z(A) \mathbb{K}(dy \times dz|x) \quad (37)$$

$$= \int_{B \times A} \mathbb{K}(dy \times dz|x) \quad (38)$$

$$= \mathbb{K}(B \times A|x) \quad (39)$$

3.8 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 3.16 (Probability set). A probability set \mathbb{P}_C 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}_C , we define marginal and conditional probabilities as probability measures and Markov kernels that satisfy Definitions 3.8 and ?? respectively for *all* base measures in \mathbb{P}_C . 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 ??) with respect to probability sets.

Definition 3.17 (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}_C , the

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

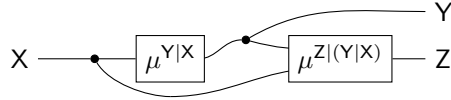
Definition 3.18 (Uniform conditional distribution). Given a sample space (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$ and a probability set \mathbb{P}_C , a uniform conditional distribution $\mathbb{P}_C^{Y|X}$ is any Markov kernel $X \rightarrow Y$ such that $\mathbb{P}_C^{Y|X}$ is an $Y|X$ conditional probability of \mathbb{P}_α for all $\mathbb{P}_\alpha \in \mathbb{P}_C$. If no such Markov kernel exists, $\mathbb{P}_C^{Y|X}$ is undefined.

Given a conditional distribution $\mu^{ZY|X}$ we can define a higher order conditional $\mu^{Z|(Y|X)}$, which is a version of $\mu^{Z|XY}$. This is useful because uniform conditionals don't always exist, but we can use higher order conditionals to show that if a probability set \mathbb{P}_C has a uniform conditional $\mathbb{P}_C^{ZY|X}$ then it also has a uniform conditional $\mathbb{P}_C^{Z|XY}$ (Theorems 6.4 and 6.5). Given $\mu^{XY|Z}$ and $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$ standard measurable, it has recently been proven that a higher order conditional $\mu^{Z|(Y|X)}$ exists Bogachev and Malofeev (2020), Theorem 3.5.

Definition 3.19 (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 $\mu^{Y|X}$,

$$\mu^{ZY|X}(B \times C|x) = \int_B \mu^{Z|(Y|X)}(C|x, y) \mu^{Y|X}(dy|x) \quad (40)$$

$$\iff \quad (41)$$



$$\mu^{ZY|X} = \quad (42)$$

Definition 3.20 (Uniform higher order conditional). 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}_C , if $\mathbb{P}_C^{ZY|X}$ exists then a uniform higher order conditional $\mathbb{P}_C^{Z|(Y|X)}$ is any Markov kernel $X \times Y \rightarrow Z$ that is a higher order conditional of some version of $\mathbb{P}_C^{ZY|X}$. If no $\mathbb{P}_C^{ZY|X}$ exists, $\mathbb{P}_C^{Z|(Y|X)}$ is undefined.

Definition 3.21 (Almost sure equality). Two Markov kernels $\mathbb{K} : X \rightarrow Y$ and $\mathbb{L} : X \rightarrow Y$ are \mathbb{P}_C, X, Y -almost surely equal if for all $A \in \mathcal{X}$, $B \in \mathcal{Y}$, $\alpha \in C$

$$\int_A \mathbb{K}(B|x) \mathbb{P}_\alpha^X(dx) = \int_A \mathbb{L}(B|x) \mathbb{P}_\alpha^X(dx) \quad (43)$$

we write this as $\mathbb{K} \stackrel{\mathbb{P}_C}{\cong} \mathbb{L}$, as the variables X and Y are clear from the context.

Equivalently, \mathbb{K} and \mathbb{L} are almost surely equal if the set $C : \{x | \exists B \in \mathcal{Y} : \mathbb{K}(B|x) \neq \mathbb{L}(B|x)\}$ has measure 0 with respect to \mathbb{P}_α^X for all $\alpha \in C$.

3.9 Extended conditional independence

Just like we defined uniform conditional probability as a version of “conditional probability” appropriate for probability sets, we need some version of “conditional independence” for probability sets. One such has already been given in some detail: it is the idea of *extended conditional independence* defined in Constantinou and Dawid (2017).

We will first define regular conditional independence. We define it in terms of a having a conditional that “ignores one of its inputs”, which, provided conditional probabilities exists, is equivalent to other common definitions (Theorem 3.23).

Definition 3.22 (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

$$\mathbb{P}^{Y|WX} \cong \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (44)$$

$$\iff \mathbb{P}^{Y|WX}(A|w, x) \cong \mathbb{P}^{\mathbb{K}(A|w)} \quad \forall A \in \mathcal{Y} \quad (45)$$

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

Theorem 3.23. *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*

$$\mathbb{P}^{Y|WX} \cong \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (46)$$

$$\iff \mathbb{P}^{Y|WX}(A|w, x) \cong \mathbb{P}^{\mathbb{K}(A|w)} \quad \forall A \in \mathcal{Y} \quad (47)$$

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

Extended conditional independence as introduced by Constantinou and Dawid (2017) allows us to define “nonstochastic variables” on the choice set C . We don’t make use of this feature here, and we limit ourselves to a special case of extended conditional independence.

Definition 3.24 (Uniform conditional independence). Given a probability set \mathbb{P}_C and variables X, Y and Z , the uniform conditional independence $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e X|Z$

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

$$\begin{array}{ccc} & Z & \text{---} \boxed{\mathbb{P}_C^{Y|Z}} \text{---} Y \\ \mathbb{P}_C^{Y|XZ} & \stackrel{\mathbb{P}_C}{\cong} & X \text{---} * \end{array} \quad (48)$$

$$\iff \quad (49)$$

$$\mathbb{P}_C^{Y|XZ}(A|x, z) \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y|Z}(A|z) \quad \forall A \in \mathcal{Y}, (x, z) \in X \times Z \quad (50)$$

Unlike the general definition of extended conditional independence, uniform conditional independence requires that the functions on the right hand side of Equation 50 are Markov kernels (Definition 3.3). Otherwise, $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e XC|Z$ is equivalent to the definition provided by Constantinou and Dawid (2017), given an appropriate choice of nonstochastic variables.

Extended conditional independence requires nonstochastic variables (here, the set “ C ”) to appear on the right hand side of the $\perp\!\!\!\perp^e$ symbol. Otherwise, for countable sets C , we can reason with collections of extended conditional independence statements as if they were regular conditional independence statements. In particular, the following rules hold:

1. Symmetry: $X \perp\!\!\!\perp_{\mathbb{P}_C}^e YC|Z$ iff $Y \perp\!\!\!\perp_{\mathbb{P}} XC|Z$
2. Decomposition: $X \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y, Z)C|W$ implies $X \perp\!\!\!\perp_{\mathbb{P}} YC|W$ and $X \perp\!\!\!\perp_{\mathbb{P}} ZC|W$
3. Weak union: $X \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y, Z)C|W$ implies $X \perp\!\!\!\perp_{\mathbb{P}_C}^e YC|(Z, W)$
4. Contraction: $X \perp\!\!\!\perp_{\mathbb{P}_C}^e ZC|W$ and $X \perp\!\!\!\perp_{\mathbb{P}_C}^e YC|(Z, W)$ implies $X \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y, Z)C|W$

3.10 Examples

Example 3.25 (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 $A \in \mathcal{Y}$, $B \in \mathcal{C}$:

$$\mathbb{P}_\alpha^{YC}(A \times B) = \mathbb{P}_\alpha(A)\delta_\alpha(B) \quad (51)$$

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

$$\mathbb{P}_\alpha^{Y|D} = \mathbb{P}_\alpha^Y \quad (52)$$

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

$$\mathbb{P}_C^{Y|D}(A|\alpha) = \mathbb{P}_\alpha^Y(A) \quad \forall A \in \mathcal{Y}, \alpha \in C \quad (53)$$

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

$$\mathbb{P}_C^{XY|Z} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \quad (54)$$

$$\iff \quad (55)$$

$$\mathbb{P}_C^{XY|Z}(A \times B|z) \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{X|Z}(A|z) \mathbb{P}_C^{Y|Z}(B|z) \quad \forall A \in \mathcal{X}, B \in \mathcal{Y}, z \in Z \quad (56)$$

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

Proof. If: By Theorem 6.5

$$\mathbb{P}_C^{XY|Z} = \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|ZX}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \quad (57)$$

$$\stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \quad (58)$$

$$= \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \quad (59)$$

Only if: Suppose

$$\mathbb{P}_C^{XY|Z} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \quad (60)$$

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

$$\mathbb{P}_C^{Y|XZ}(B|x, z) > \mathbb{P}_C^{Y|Z}(B|z) \quad \forall (x, z) \in A \times C \quad (61)$$

then

$$\mathbb{P}_\alpha^{\mathbf{X}^{\mathbf{Y}\mathbf{Z}\mathbf{Z}}}(A \times B \times C) = \int_{A \times C} \mathbb{P}_C^{\mathbf{Y}|\mathbf{X}\mathbf{Z}}(B|x, z) \mathbb{P}_C^{\mathbf{X}|\mathbf{Z}}(dx|z) \mathbb{P}_\alpha^{\mathbf{Z}}(dz) \quad (62)$$

$$> \int_{A \times C} \mathbb{P}_C^{\mathbf{Y}|\mathbf{X}}(B|z) \mathbb{P}_C^{\mathbf{X}|\mathbf{Z}}(dx|z) \mathbb{P}_\alpha^{\mathbf{Z}}(dz) \quad (63)$$

$$= \int_C \mathbb{P}_C^{\mathbf{X}^{\mathbf{Y}|\mathbf{X}}}(A \times B|z) \mathbb{P}_\alpha^{\mathbf{Z}}(dz) \quad (64)$$

$$= \mathbb{P}_\alpha^{\mathbf{X}^{\mathbf{Y}\mathbf{Z}\mathbf{Z}}}(A \times B \times C) \quad (65)$$

a contradiction. An analogous argument follows if we replace “>” with “<” in Eq. 61. \square

4 When do response functions exist?

We model decision problems with probability sets \mathbb{P}_C for some set of choices C . If we have a pair of variables \mathbf{X} and \mathbf{Y} such that the uniform conditional $\mathbb{P}_C^{\mathbf{Y}|\mathbf{X}}$ exists (Definition 3.18), then the joint outcome $\mathbb{P}_\alpha^{\mathbf{X}\mathbf{Y}}$ of any choice $\alpha \in C$ can be computed from the marginal distribution $\mathbb{P}_\alpha^{\mathbf{X}}$ alone.

We’re interested in models that feature a particular kind of “causal effect” that we call a *conditionally independent and identical response functions*, or just “response functions” for short. They are a causal analogue of conditionally independent and identical sequences of random variables. Concretely, a model with response functions is a probability set \mathbb{P}_C with variables $\mathbf{Y} := (\mathbf{Y}_i)_{i \in M}$, $(\mathbf{X}_i)_{i \in M}$ for some index set M and some \mathbf{H} such that $\mathbf{Y} \perp\!\!\!\perp_{\mathbb{P}_C}^e C | \mathbf{X}_i \mathbf{H}$ and $\mathbf{H} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{X}_i C$ (see Section ?? for extended conditional independence) and $\mathbb{P}_C^{\mathbf{Y}_i | \mathbf{X}_i \mathbf{H}} = \mathbb{P}_C^{\mathbf{Y}_j | \mathbf{X}_j \mathbf{H}}$ for all $i, j \in M$ (the identical response conditional requirement).

is that OK?

We will focus on the case where $\mathbb{P}_\alpha^{\mathbf{H} | \mathbf{Y}_A \mathbf{X}_A}$ approaches a deterministic distribution as $|A| \rightarrow \infty$, for appropriate $\alpha \in C$. We could say this is the case of “identifiable” response conditionals.

Put together, these conditions say: in the limit of infinite samples under an appropriate sampling regime $\alpha \in C$, the model converges to a probabilistic function $X \rightarrow Y$ that represents “the probability of \mathbf{Y}_i given \mathbf{X}_i ” for any unobserved $(\mathbf{X}_i, \mathbf{Y}_i)$. For arbitrary $\alpha' \in C$, we may instead converge to a set containing this limiting distribution. We think – although it usually isn’t stated in these terms – that given a causal Bayesian network, if the function “ $x \mapsto \mathbb{P}(\mathbf{Y} | \text{do}(\mathbf{X} = x))$ ” is identifiable, then it is an instance of the kind of function that we described in the previous sentences.

We prove our result under the simplifying assumption that $\mathbf{X}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{Y}_{<i} C | \mathbf{X}_{<i}$. This is a limiting assumption – for example, it excludes cases where \mathbf{X}_i depends on $(\mathbf{X}_{<i}, \mathbf{Y}_{<i})$. In the more general case where this does not hold, the conditions we provide must still hold for any $C' \subset C$ such that $\mathbf{X}_i \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{Y}_{<i>C' | \mathbf{X}_{<i}$, but providing sufficient conditions in this case is the topic of further work.

Under this assumption, we show that *causal contractibility* is necessary and sufficient for the existence of response conditionals. Causal contractibility can

be broken down into two sub-assumptions: *exchange commutativity* and *consequence locality*. The first is the assumption that the uniform conditional probability $\mathbb{P}_C^{Y|X}$ “commutes” with the permutation operation, and the second is the assumption that X_i “has no effect” on any Y_j for $j \neq i$.

4.1 Relevance to previous work

Both sub-assumptions have precedent in existing literature, but these precedents tend to have been stated at somewhat informally.

Post-treatment exchangeability found in Dawid (2021) is implied by exchange commutativity, but not the reverse. “Causal exchangeability” notions are also 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 given as symmetries of *decision procedures* – they involve actually swapping actions taken or individuals in an experiment – while exchange commutativity is a symmetry of probability sets.

Consequence locality is similar to the stable unit treatment distribution assumption (SUTDA) in Dawid (2021), although consequence locality is distinguished by being a concrete extended conditional independence (Definition 4.1) while SUTDA is given as the assumption that Y_i “depends only on” X_i . Consequence locality is also similar to 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.

String diagram statements of both sub-assumptions can give some intuition about what they mean. For clarity, these diagrams illustrate the assumptions with exactly two inputs and two outputs, while the general definitions are for any countable number of inputs and outputs.

Exchange commutativity for two inputs and outputs is given by the following equality:

$$\begin{array}{c} D_1 \\ D_2 \end{array} \begin{array}{c} \diagup \\ \diagdown \end{array} \boxed{\mathbb{P}_C^{Y_{\{1,2\}}|D_{\{1,2\}}}} \begin{array}{c} \diagdown \\ \diagup \end{array} \begin{array}{c} Y_1 \\ Y_2 \end{array} = \begin{array}{c} D_1 \\ D_2 \end{array} \boxed{\mathbb{P}_C^{Y_{\{1,2\}}|D_{\{1,2\}}}} \begin{array}{c} \diagdown \\ \diagup \end{array} \begin{array}{c} Y_1 \\ Y_2 \end{array} \quad (66)$$

While consequence locality for two inputs and outputs is given by the following pair of equalities:

$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{array}{|c|} \hline \mathbb{P}_C^{Y_{1,2}|X_{1,2}} \\ \hline \end{array} \begin{array}{c} Y_1 \\ * \end{array} = \begin{array}{c} X_1 \\ X_2 \end{array} \begin{array}{|c|} \hline \mathbb{P}_C^{Y_1|X_1} \\ \hline \end{array} \begin{array}{c} Y_1 \\ * \end{array} \quad (67)$$

$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{array}{|c|} \hline \mathbb{P}_C^{Y_{1,2}|X_{1,2}} \\ \hline \end{array} \begin{array}{c} * \\ Y_2 \end{array} = \begin{array}{c} X_1 \\ X_2 \end{array} \begin{array}{|c|} \hline \mathbb{P}_C^{Y_2|X_2} \\ \hline \end{array} \begin{array}{c} * \\ Y_2 \end{array} \quad (68)$$

4.2 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 4.1 (Locality of consequences). Suppose we have a sample space (Ω, \mathcal{F}) and a probability set \mathbb{P}_C 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 $A \subset M$, $\mathbf{Y}_A \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{D}_{A^c} C | \mathbf{D}_A$ then \mathbb{P}_C 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.

Note that consequence locality implies $\mathbf{Y}_M \perp\!\!\!\perp_{\mathbb{P}_C}^e C | \mathbf{D}_M$, and hence we have the uniform conditional $\mathbb{P}_C^{\mathbf{Y}_M | \mathbf{D}_M}$. We assume the existence of such a conditional for the next definition.

Definition 4.2 (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 4.3 (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 $\rho : M \rightarrow M$

$$\text{swap}_{\rho(\mathbf{D})} \mathbb{P}_C^{\mathbf{Y} | \mathbf{D}} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{\mathbf{Y} | \mathbf{D}} \text{swap}_{\rho(\mathbf{Y})} \quad (69)$$

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

Theorem 4.4 shows that neither condition implies the other.

Theorem 4.4. *Exchange commutativity does not imply locality of consequences or vice versa.*

Proof. Appendix 6.6. □

If we are modelling the treatment of several patients whom who have already been examined, we might assume consequence locality – patient B’s treatment does not affect patient A – but not exchange commutativity – we don’t expect the same results from giving patient A’s treatment to patient B as we would from giving patient A’s treatment to patient A.

A model of stimulus payments might exhibit exchange commutativity but not consequence locality. If exactly n payments of \$10 000 are made, we might suppose that it doesn’t matter much exactly who receives the payments, but the amount of inflation induced depends on the number of payments made; making 100 such payments will have a negligible effect on inflation, while making payments to everyone in the country will have a substantial effect. Dawid (2000) offers the example of herd immunity in vaccination campaigns as a situation where post-treatment exchangeability holds but locality of consequences does not.

Although locality of consequences seems to intuitively encompass an assumption of non-interference, it still allows for some models in which exhibit certain kinds of interference between actions and outcomes of different indices. For example: I have an experiment where I first flip a coin and record the results of this flip as the outcome of the first step of the experiment, but I can choose either to record this same outcome as the provisional result of the second step (this is the choice $D_1 = 0$), or choose to flip a second coin and record the result of that as the provisional result of the second step of the experiment (this is the choice $D_1 = 1$). At the second step, I may further choose to copy the provisional results ($D_2 = 0$) or invert them ($D_2 = 1$). Then

$$\mathbb{P}_S^{Y_1|D}(y_1|d_1, d_2) = 0.5 \quad (70)$$

$$\mathbb{P}_S^{Y_2|D}(y_2|d_1, d_2) = 0.5 \quad (71)$$

- The marginal distribution of both experiments in isolation is Bernoulli(0.5) no matter what choices I make, so a model of this experiment would satisfies Definition 4.1
- 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 4.5 (Causal contractibility). A probability set \mathbb{P}_C is $(D; Y)$ -*causally contractible* if it is both exchange commutative and exhibits consequence locality.

Theorem 4.6 (Equality of reduced conditionals). *A probability set \mathbb{P}_C that is $(D; Y)$ -causally contractible has, for any $A, B \subset M$ with $|A| = |B|$*

$$\mathbb{P}_C^{Y_A|D_A} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_B|D_B} \quad (72)$$

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.

$$\begin{array}{c} D_A \text{---} \boxed{\mathbb{P}_C^{Y_A|D_A}} \text{---} Y_A \\ D_{M \setminus A} \text{---} * \end{array} = \begin{array}{c} D_{M \setminus A} \text{---} \boxed{\mathbb{P}_C^{Y_A Y_{M \setminus A} | D_A D_{M \setminus A}}} \text{---} Y_A \\ D_{M \setminus A} \text{---} * \end{array} \quad (73)$$

$$= \begin{array}{c} D_{M \setminus A} \text{---} \boxed{s_{BA}} \text{---} \boxed{\mathbb{P}_C^{Y_A Y_{M \setminus A} | D_A D_{M \setminus A}}} \text{---} \boxed{s_{AB}} \text{---} Y_A \\ D_{M \setminus A} \text{---} * \end{array} \quad (74)$$

$$= \begin{array}{c} D_{M \setminus B} \text{---} \boxed{\mathbb{P}_C^{Y_B Y_{M \setminus B} | D_A D_{M \setminus B}}} \text{---} Y_B \\ D_{M \setminus B} \text{---} * \end{array} \quad (75)$$

Thus

$$\mathbb{P}_C^{Y_A | D_A D_{M \setminus A}} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_B | D_B D_{M \setminus B}} \quad (76)$$

$$\stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} D_A \text{---} \boxed{\mathbb{P}_C^{Y_A | D_A}} \text{---} Y_A \\ D_{M \setminus A} \text{---} * \end{array} \quad (77)$$

$$\Rightarrow \mathbb{P}_C^{Y_A | D_A} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_B | D_B} \quad (78)$$

□

4.3 Existence of response conditionals

The main result in this section is Theorem 4.9 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 $\mathbb{P}_C^{Y_1 | D_1 H}$. Note the hypothesis H that appears in this conditional; it can be given the interpretation of a random variable that expresses the “true but initially unknown” $Y_1 | D_1$ conditional probability.

Theorem 4.7. *Given a probability set \mathbb{P}_C and variables $D := (D_i)_{i \in \mathbb{N}}$ 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*

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \triangle \mu^{Y^D} \\ D \text{---} \triangle \text{---} \boxed{\mathbb{F}_{\text{ev}}} \text{---} Y \end{array} \quad (79)$$

$$\iff \quad (80)$$

$$\mathbb{P}_C^{Y|D}(y | (d_i)_{i \in \mathbb{N}}) = \mu^{Y^D} \Pi_{(d_i i)_{i \in \mathbb{N}}}(y) \quad (81)$$

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

$$ev : D^{\mathbb{N}} \times Y^{D \times \mathbb{N}} \rightarrow Y \quad (82)$$

$$((d_i)_{i \in \mathbb{N}}, (y_{ij})_{i \in D, j \in \mathbb{N}}) \mapsto (y_{d_i i})_{i \in \mathbb{N}} \quad (83)$$

Proof. Appendix 6.6. \square

We would prefer to talk about Y^D as a latent variable, rather than needing to refer to the factorisation of a model in terms of μ^{Y^D} in Equation 79. This motivates the definition of an *augmented* causally contractible model.

Lemma 4.8 (Augmented causally contractible model). *Given a $(D; Y)$ -causally contractible model \mathbb{P}'_C on (Ω', \mathcal{F}') , there exists an augmented model \mathbb{P}_C on $(\Omega, \mathcal{F}) := ((\Omega' \times Y^D, \mathcal{F}' \otimes \mathcal{Y}^D)$ such that $\mathbb{P}_C \Pi_{\Omega'} = \mathbb{P}'_C$ and, defining $Y^D : \Omega \times Y^D \rightarrow Y^D$ as the projection onto Y^D*

$$\mathbb{P}_C^{Y^D} = \begin{array}{c} \text{Diagram: A triangle labeled } \mathbb{P}_C^{Y^D} \text{ with } D \text{ at the bottom vertex and } Y^D \text{ at the top vertex. A line connects } D \text{ to a box labeled } \mathbb{F}_{ev}, \text{ which then connects to } Y. \end{array} \quad (84)$$

Proof. Appendix 6.6. \square

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 , and the value of Y_i given D is deterministically equal to the Y_i^D (abusing notation). 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 4.7 yields $\mathbb{P}^{Y^D}(y_i^D) = \prod_{j \in D} 0.5$ for all $y_i^D \in Y^D$. In this model Y_i^0 and Y_i^1 are independent and uniformly distributed. However, if we wanted Y_i^0 to be interpretable as “what would happen if I bet on outcome 0 on turn i ” and Y^1 to represent “what would happen if I bet on outcome 1 on turn i ”, then we ought to have $Y_i^0 = 1 - Y_i^1$.

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

Theorem 4.9. Suppose we have a sample space (Ω, \mathcal{F}) and a probability set \mathbb{P}_C and variables $\mathbf{D} := (D_i)_{i \in \mathbb{N}}$ and $\mathbf{Y} := (Y_i)_{i \in \mathbb{N}}$. Suppose also that \mathbb{P}_C is $(\mathbf{D}; \mathbf{Y})$ -causally contractible if and only if there exists some $\mathbf{H} : \Omega \rightarrow H$ such that $\mathbb{P}_C^{\mathbf{H}}$ and $\mathbb{P}_C^{Y_i | \mathbf{H} D_i}$ exist for all $i \in \mathbb{N}$ and

$$\mathbb{P}_C^{Y | \mathbf{D}} = \begin{array}{c} \triangle^{\mathbf{H}} \\ \mu^{\mathbf{H}} \\ \text{---} \end{array} \begin{array}{c} \text{---} \end{array} \begin{array}{c} \boxed{\Pi_{D,i}} \boxed{\mathbb{P}_C^{Y_0 | \mathbf{H} D_0}} \text{---} Y_i \\ i \in \mathbb{N} \end{array} \quad (85)$$

$$\iff \quad (86)$$

$$Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e Y_{\mathbb{N} \setminus i}, D_{\mathbb{N} \setminus i} C | \mathbf{H} D_i \quad \forall i \in \mathbb{N} \quad (87)$$

$$\wedge \mathbf{H} \perp\!\!\!\perp_{\mathbb{P}_C}^e D C \quad (88)$$

$$\wedge D_i \perp\!\!\!\perp_{\mathbb{P}_C}^e Y_{<i>} C | D_{<i>} \quad (89)$$

$$\wedge \mathbb{P}_C^{Y_i | \mathbf{H} D_i} = \mathbb{P}^{Y_0 | \mathbf{H} D_0} \quad \forall i \in \mathbb{N} \quad (90)$$

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

Proof. Appendix 6.6. \square

4.4 Elaborations and examples

Theorem 4.9 requires an infinite sequence of causally contractible pairs. In practice we only want to model finite sequences of variables, but this theorem applies as long as it is possible to extend the finite model to an infinite model maintaining causal contractibility.

Theorem 4.9 applies whatever procedure we use to obtain the (D_i, Y_i) pairs – the D_i s may be randomised, passive observations or active choices. Purely passive observations can be modeled with a probability set of size 1, and in this case an exchangeable sequence of (D_i, Y_i) will also be causally contractible.

If we are modelling M passive observations followed by N active choices, then we will have a model \mathbb{P}_C with $D_{[M]} \perp\!\!\!\perp_{\mathbb{P}_C}^e C$ (because these are passive observations). If this model is $(\mathbf{D}; \mathbf{Y})$ -causally contractible, then one consequence of this is an “observational imitation” condition: any choice α that makes $\mathbb{P}_\alpha^{D_{[M+N]}}$ exchangeable also makes $\mathbb{P}_\alpha^{Y_{[M+N]}}$ exchangeable. That is, if for some permutation swap_ρ

$$\mathbb{P}_\alpha^{D_{[M+N]}} \text{swap}_\rho = \mathbb{P}_\alpha^{D_{[M+N]}} \quad (91)$$

then by commutativity of exchange

$$\mathbb{P}_\alpha^{Y_{[M+N]}} = \mathbb{P}_\alpha^{D_{[M+N]}} \mathbb{P}_C^{Y_{[M+N]} | D_{[M+N]}} \quad (92)$$

$$= \mathbb{P}_\alpha^{D_{[M+N]}} \text{swap}_\rho \mathbb{P}_C^{Y_{[M+N]} | D_{[M+N]}} \quad (93)$$

$$= \mathbb{P}_\alpha^{D_{[M+N]}} \mathbb{P}_C^{Y_{[M+N]} | D_{[M+N]}} \text{swap}_\rho \quad (94)$$

$$= \mathbb{P}_\alpha^{Y_{[M+N]}} \text{swap}_\rho \quad (95)$$

If we assume a probability set \mathbb{P}_C is $(D, X; Y)$ -causally contractible and $X_i \perp\!\!\!\perp_{\mathbb{P}_C}^e D_i C | H$ – that is, D_i is must be independent of X_i conditional on H – then we get a version of the “backdoor adjustment” formula. Specifically

$$\mathbb{P}_\alpha^{Y_i | D_i H}(A | d, h) = \int_X \mathbb{P}_\alpha^{Y_i | X_i D_i H}(A | d, x, h) \mathbb{P}_\alpha^{X_i | D_i H}(dx | d, h) \quad (96)$$

$$= \int_X \mathbb{P}_C^{Y_i | X_i D_i H}(A | d, x, h) \mathbb{P}_C^{X_i | H}(dx | h) \quad (97)$$

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

$$\mathbb{P}_\alpha^{Y_i | D_i H}(A | d, h) = \int_X \mathbb{P}_C^{Y_i | X_i D_i H}(A | d, x, h) \mathbb{P}_C^{X_1 | H}(dx | h) \quad (98)$$

Equation 98 is identical to the backdoor adjustment formula for an intervention on D_1 targeting Y_1 where X_1 is a common cause of both.

While it is formally possible to use a causally contractible model for a decision procedure that involves both passive observations and active choices, causal contractibility is a very strong assumption. Suppose we have a decision procedure in which M passive observations are made (D_M, Y_M) , followed by M active choices $(D_{(M, 2M]}, Y_{(M, 2M]})$. If a model \mathbb{P}_C of this procedure is (D_{2M}, Y_{2M}) -causally contractible model then the following holds (see corollary 4.6):

$$\mathbb{P}_C^{Y_{[2, M+1]} | D_{[2, M+1]}} = \mathbb{P}^{Y_{(M, 2M]} | D_{(M, 2M]}} \quad (99)$$

$$\implies \mathbb{P}_C^{Y_{M+1} | D_{[2, M+1]} Y_{[2, M]}} = \mathbb{P}^{Y_{M+1} | D_{(M, 2M]} Y_{(M+1, 2M]}} \quad (100)$$

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. Normally one might consider randomised experimental results to be “better” than passive observations, but this is not compatible with the assumption of causal contractibility.

4.5 Assessing causal contractibility

Assessing when a particular sequence of experiments should be modeled with a causally contractible model can be difficult. One way to justify the assumption is in two steps: first, all the repetitions of the experiment that yield the values of each of the (D_i, Y_i) pairs are indistinguishable “at the time of model construction”, and and they are still indistinguishable after learning the value of D – because, for example, D is deterministic for each choice $\alpha \in C$.

Two step justifications of this form are common in literature on causal identifiability. 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.

Note that exchanging individuals involved in an experiment and exchanging the individuals' exposure states are two different things, and the former doesn't imply the latter. We may consider a model that is symmetric to permutations of individual identifiers but is not symmetric to permuting individual identifiers and leaving exposure states fixed.

Dawid (2021) suggests (with many qualifications) that “post-treatment exchangeability” for a decision problem regarding taking aspirin to treat a headache may be acceptable if the data are from

A group of individuals whom I can regard, in an intuitive sense, as similar to myself, with headaches similar to my own.

Dawid points to the “first step” in our two step justification for causal contractibility: that the people involved are “similar” in an appropriate sense. However, under Dawid's approach there is a background assumption here that whether or not I take the aspirin is deterministic given the choice I end up making, which is the second step in our two step justification.

Finally, Rubin (2005) explicitly discusses two separate assumptions to justify causal identifiability:

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}

Here we have a more abstract statement about the row-exchangeability of “the science”, rather than individual people involved in an experiment, but we regard it as similar in spirit to assumptions that people involved in the experiment are “similar”. Rubin explicitly mentions a second condition: that the treatment assignment is randomized.

Theorem 4.13 formalises these ideas. As an example of its application, consider an experiment where N patients, each with an individual identifier I_i , receive treatment D_i and experience outcome Y_i . We assume a $((D, I); Y)$ -causally contractible model \mathbb{P}_C is appropriate. This reflects two judgments; firstly, that treatment D_i and identifiers I_i screen off all other variables from Y_i (Definition 4.1), and secondly that the order in which the individuals appear and the treatments are received does not alter the consequences we expect to see (Definition 4.3). The fact that we need a preliminary assumption of causal contractibility is similar to how, in the potential outcomes framework, a preliminary assumption of SUTVA is required in order to justify the use of potential outcomes.

Next, we assume that, no matter which choice $\alpha \in C$ is decided on, all identifiers can be swapped without altering the distribution over consequences, and finally that for each choice $\alpha \in C$ the treatment vector D is deterministic. Then, according to Theorem 4.13, \mathbb{P}_C is also $(D; Y)$ -causally contractible. This can be extended to the case where D is a function of a “random signal” R .

Lemma 4.10. *If \mathbb{P}_C is $((D, I); Y)$ -causally contractible and $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e I \mid D$, then \mathbb{P}_C is also $(D; Y)$ -causally contractible.*

Proof. For arbitrary $\nu \in \Delta(I^{\mathbb{N}})$, by assumption of $((D, I); Y)$ -causal contractibility and Theorem 4.9

$$\mathbb{P}_C^{\mathbf{Y}|\mathbf{D}\mathbf{I}} = \begin{array}{c} \text{D} \\ \text{I} \end{array} \begin{array}{c} \text{---} \text{ } \Pi_{D,i} \\ \text{---} \text{ } \Pi_{I,i} \end{array} \begin{array}{c} \text{---} \text{ } \mathbb{P}_C^{\mathbf{Y}_1|\mathbf{H}\mathbf{D}_1\mathbf{I}_1} \\ \text{---} \text{ } \mathbf{Y}_i \end{array} \quad (101)$$

$$= \text{D} \begin{array}{c} \text{---} \bullet \text{---} \text{P}_{D,i} \\ \text{---} \bullet \text{---} \text{P}_{I,i} \end{array} \begin{array}{c} \text{---} \text{P}_C^{\mathbf{Y}_i | \mathbf{H}_{D1} \mathbf{I}_1} \text{---} \mathbf{Y}_i \end{array} \quad (102)$$

$$= \text{D} \begin{array}{c} \text{I} \rightarrow * \end{array} \left(\begin{array}{c} \text{P}_C^H \\ \text{P}_C^{Y_1|H D_1 I_1} \\ \text{P}_1 \end{array} \right) \text{Y}_i \quad (103)$$

$$\Rightarrow \mathbb{P}_C^{Y|D} = \begin{array}{c} \begin{array}{c} \text{Diagram of } \mathbb{P}_C^{Y|D} \text{ block} \end{array} \end{array} \quad (104)$$

Applying Theorem 4.9 in reverse, we get \mathbb{P}_C is $(D; Y)$ -causally contractible. \square

Definition 4.11 (Index variable). Suppose we have a probability set \mathbb{P}_C and a variable $\mathbf{l} : \Omega \rightarrow \mathbb{N}^{\mathbb{N}}$, such that, defining $A \subset \mathcal{P}(\mathbb{N})$ to be the set of all permutations of \mathbb{N} , $\mathbb{P}_\alpha^{\mathbf{l}}(A) = 1$. Then \mathbf{l} is an *index variable*.

Lemma 4.12. *Suppose we have a probability set \mathbb{P}_C where $\mathbf{Y} \perp\!\!\!\perp_{\mathbb{P}_C}^e C | (\mathbf{D}, \mathbf{I})$ and \mathbf{I} is an index variable. If for each permutation $\rho : \mathbb{N} \rightarrow \mathbb{N}$*

$$\mathbb{P}_\alpha^{\mathbf{Y}|\mathbf{ID}} = (\text{swap}_{\rho(I)} \otimes Id_X) \mathbb{P}_\alpha^{\mathbf{Y}|\mathbf{ID}} \quad (105)$$

then $Y \perp\!\!\!\perp^e_{\mathbb{P}_C} IC|D$.

Proof. Taking $A \subset \mathbb{N}^{\mathbb{N}}$ to be the set of permutations of \mathbb{N} , note that for every $i \in A$, $B \in \mathcal{Y}^{\mathbb{N}}$, $d \in D^{\mathbb{N}}$ we can take $\rho_i : \mathbb{N} \rightarrow \mathbb{N}$ such that the image of i under

ρ is \mathbb{N} ; that is, $\rho_i(i) = \text{id}_{\mathbb{N}}$. Then

$$\mathbb{P}_C^{Y|\text{ID}}(B|i, d) = (\text{swap}_{\rho_i(I)} \otimes \text{Id}_X) \mathbb{P}_C^{Y|\text{ID}}(B|i, d) \quad (106)$$

$$= \mathbb{P}_C^{Y|\text{ID}}(B|\text{id}_{\mathbb{N}}, d) \quad (107)$$

Therefore

$$\mathbb{P}_C^{Y|\text{ID}} \stackrel{\mathbb{P}_C}{\cong} \text{erase}_{\mathbb{N}^{\mathbb{N}}} \otimes \mathbb{K} \quad (108)$$

where $\mathbb{K} : D^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is the kernel

$$(B|d) \mapsto \mathbb{P}_C^{Y|\text{ID}}(B|\text{id}_{\mathbb{N}}, d) \quad (109)$$

□

Theorem 4.13. *Suppose we have a probability set \mathbb{P}_C , C countable, that $((D, \text{I}); Y)$ -causally contractible for variables $Y : \Omega \rightarrow Y^{\mathbb{N}}$, $D : \Omega \rightarrow D^{\mathbb{N}}$ and index variable $\text{I} : \Omega \rightarrow \mathbb{N}^{\mathbb{N}}$. If for each $\alpha \in C$, $\rho : \mathbb{N} \rightarrow \mathbb{N}$*

$$\mathbb{P}_{\alpha}^{Y|\text{I}} = \text{swap}_{\rho(I)} \mathbb{P}_{\alpha}^{Y|\text{I}} \quad (110)$$

and there is an invertible function $f : C \rightarrow D$ such that

$$\mathbb{P}_{\alpha}^D = \mathbb{F}_f \quad (111)$$

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

Proof. The map $\mathbb{Q} : (B|i, \alpha) \mapsto \mathbb{P}_{\alpha}^{Y|\text{I}}(B|i)$ is itself a Markov kernel. By assumption, for any α ,

$$\mathbb{Q} \stackrel{\mathbb{P}_{\alpha}}{\cong} (\text{id}_{\mathbb{N}^{\mathbb{N}}} \otimes \mathbb{F}_f) \mathbb{P}_{\alpha}^{Y|\text{ID}} \quad (112)$$

Furthermore, by assumption

$$(\text{swap}_{\rho(I)} \otimes \text{Id}_C) \mathbb{Q} = \mathbb{Q} \quad (113)$$

Therefore

$$(\text{swap}_{\rho(I)} \otimes \text{Id}_C) \mathbb{P}_{\alpha}^{Y|\text{ID}} = (\text{id}_{\mathbb{N}^{\mathbb{N}}} \otimes \mathbb{F}_f)(\text{id}_{\mathbb{N}^{\mathbb{N}}} \otimes \mathbb{F}_{f^{-1}})(\text{swap}_{\rho(I)} \otimes \text{Id}_C) \mathbb{Q} \quad (114)$$

$$= (\text{id}_{\mathbb{N}^{\mathbb{N}}} \otimes \mathbb{F}_{f^{-1}}) \mathbb{Q} \quad (115)$$

$$= \mathbb{P}_{\alpha}^{Y|\text{ID}} \quad (116)$$

Then by Lemma 4.12 we have $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e \text{I}C|D$ and by Lemma 4.10 we have $(D; Y)$ -causal contractibility. □

If we suppose Theorem 4.13 applies to $C' \subset C$ such that D is deterministic for all $\alpha \in C'$, while C consists of C' and all “random choices between elements of C' ”; that is for all $\beta \in C$

$$\mathbb{P}_\beta = \sum_{c \in C} k_c \mathbb{P}_c \quad (117)$$

Then it follows that

$$\mathbb{P}_\beta^{Y|D} = \sum_{c \in C} k_c \mathbb{P}_c^{Y|D} \quad (118)$$

and hence \mathbb{P}_C is still $(D; Y)$ -causally contractible. Note that in order to actually implemented a random choice, we would typically consult a known random source R and set D deterministically on the basis of the value of R .

4.6 Body mass index revisited

If we have a probability set \mathbb{P}_C with $B := (B_i)_{i \in M}$ representing body mass index and $Y := (Y_i)_{i \in M}$ representing health outcomes of interest, the previous considerations don't support a judgement of causal contractibility for $(B; Y)$, because the choices we imagine we might have available do not allow B to be a deterministic invertible function of the choice. Note that we haven't established that causal contractibility cannot be appropriate, merely that we have no reason to accept it on the basis of arguments so far.

Causal contractibility is the a priori assumption that there is a response function relating each pair from a sequence of pairs of variables. However, we could also consider the possibility that we conclude that there is such a response function after reviewing the data.

Suppose we are in possession of a $(D; (B, Y))$ -causally contractible probability set \mathbb{P}_C , such that each (D_i, B_i, Y_i) is related by the response conditional $\mathbb{P}_C^{Y_i B_i | D_i H}$. Suppose that we also have an “oracle” available that performs an infinite number of samples under appropriate conditions and reveals the value $h \in H$ yielded by the variable H . Then we can consider the new probability set $\mathbb{P}_{C,h}$ where for arbitrary $Z : \Omega \rightarrow Z$, $A \in \mathcal{Z}$

$$\mathbb{P}_{C,h}^Z(A) = \mathbb{P}_C^{Z|H}(A|h) \quad (119)$$

Note that $\mathbb{P}_{C,h}$ remains $(D; (B, Y))$ -causally contractible with response conditionals $\mathbb{P}_{C,h}^{Y_i B_i | D_i}$. Furthermore that by Theorem 6.5, we have $((D, B); Y)$ -causal contractibility with response conditionals $\mathbb{P}_{C,h}^{Y_i | D_i B_i}$. In this case, we could find that $Y_i \perp\!\!\!\perp_{\mathbb{P}_{C,h}}^e D_i | B_i$, and so by Lemma 4.10 $\mathbb{P}_{C,h}$ is also $(B; Y)$ -causally contractible.

In this case, it seems much more reasonable to describe the fact that B has a causal effect on Y as a *finding*, rather than an *assumption*.

5 Conclusion

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

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

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

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

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

A property of conditional probability models that may help bridge this gap

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

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

5.1 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.

6 Appendix, needs to be organised

6.1 Markov categories

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

Definition 6.1. 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

$$\text{del}_X := \text{---} * \text{copy}_X \quad := \text{---} \text{---} \text{---} \quad (120)$$

and satisfying the commutative comonoid equations

$$\text{---} \text{---} \text{---} = \text{---} \text{---} \text{---} \quad (121)$$

$$\text{---} \text{---} * = \text{---} = \text{---} * \quad (122)$$

$$\text{---} \text{---} = \text{---} \text{---} \quad (123)$$

as well as compatibility with the monoidal structure

$$\begin{array}{ccc} X \otimes Y & \longrightarrow & * \\ & & \downarrow \\ & & X \longrightarrow * \end{array} \quad (124)$$

$$X \otimes Y \longrightarrow \begin{array}{c} \text{---} \curvearrowright X \otimes Y \\ \text{---} \curvearrowleft X \otimes Y \end{array} = \begin{array}{c} X \text{---} \bullet \text{---} X \\ Y \text{---} \bullet \text{---} Y \end{array} \begin{array}{c} \text{---} \curvearrowright X \\ \text{---} \curvearrowleft Y \end{array} \quad (125)$$

and the naturality of del , which means that

$$\text{---} \boxed{f} \text{---} * \text{---} * = \text{---} * \text{---} * \quad (126)$$

for every morphism f .

6.2 Existence of conditional probabilities

Lemma 6.2 (Conditional pushforward). *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}_C with conditional $\mathbb{P}_C^{\mathbf{X}|\mathbf{Y}}$ such that $\mathbf{Z} = f \circ \mathbf{Y}$ for some $f : Y \rightarrow Z$. Then there exists a conditional probability $\mathbb{P}_C^{\mathbf{Z}|\mathbf{X}} = \mathbb{P}_C^{\mathbf{Y}|\mathbf{X}} \mathbb{F}_f$.*

Proof. Note that $(X, Z) = (\text{id}_X \otimes f) \circ (X, Y)$. Thus, by Lemma 3.11, for any $\mathbb{P}_\alpha \in \mathbb{P}_C$

$$\mathbb{P}_\alpha^{\mathbf{XZ}} = \mathbb{P}_\alpha^{\mathbf{XY}} \mathbb{F}_{\text{id}_X \otimes f} \quad (127)$$

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

$$\mathbb{F}_{\text{id}_X \otimes f}(A \times B|x, y) = \delta_x(A)\delta_{f(y)}(B) \quad (128)$$

$$= \mathbb{F}_{\text{id}_X}(A|x) \otimes \mathbb{F}_f(B|y) \quad (129)$$

$$\implies \mathbb{F}_{\text{id}_X \otimes f} = \mathbb{F}_{\text{id}_X} \otimes \mathbb{F}_f \quad (130)$$

Thus

$$\mathbb{P}_\alpha^{\mathbf{XZ}} = (\mathbb{P}_\alpha^{\mathbf{X}} \odot \mathbb{P}_C^{\mathbf{Y|X}}) \mathbb{F}_{\text{id}_X} \otimes \mathbb{F}_f \quad (131)$$

$$= \text{Diagram (132)} \quad (132)$$

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

Theorem 6.3 (Existence of regular conditionals). *Suppose we have a sample space (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$ with Y standard measurable and a probability model \mathbb{P}_α on (Ω, \mathcal{F}) . Then there exists a conditional $\mathbb{P}_\alpha^{Y|X}$.*

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

Theorem 6.4 (Existence of higher order valid conditionals with respect to probability sets). *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}_C with regular conditional $\mathbb{P}_C^{YZ|X}$ and Y and Z standard measurable. Then there exists a regular $\mathbb{P}_C^{Z|(Y|X)}$.*

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 $f \circ (Y, Z) = Y$. 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

$$n(f^{-1}(y)|x, y) = 1 \quad (133)$$

$$m(Y^{-1}(A) \cap B|x) = \int_A n(B|x, y) m\mathbb{F}_f(dy|x) \forall A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z} \quad (134)$$

In particular, $\mathbb{P}_C^{YZ|X}$ is a Borel measurable map $X \rightarrow Y \times Z$. Thus equation 134 implies for all $A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z}$

$$\mathbb{P}_C^{YZ|X}(Y^{-1}(A) \cap B|x) = \int_A n(B|x, y) \mathbb{P}_C^{YZ|X} \mathbb{F}_f(dy|x) \quad (135)$$

$$= \int_A n(B|x, y) \mathbb{P}_C^{Y|X}(dy|x) \quad (136)$$

Where Equation 136 follows from Lemma 6.2.

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

$$\mathbb{P}_C^{YZ|X}(Y^{-1}(A) \cap B|x) = \int_A n(B|x, y) \mathbb{P}_\alpha^{Y|X}(dy|x) \quad (137)$$

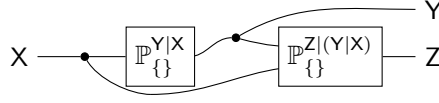
which implies n is a version of $\mathbb{P}_C^{Z|(Y|X)}$. By Lemma 6.2, $n\mathbb{F}_f$ is a version of $\mathbb{P}_C^{Z|(Y|X)}$. \square

We might be motivated to ask whether the higher order conditionals in Theorem 6.4 can be chosen to be valid. Despite Lemma 6.9 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 133 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 6.5 (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}_C with conditional $\mathbb{P}_C^{YZ|X}$. Then $\mathbb{P}_C^{Z|(Y|X)}$ is a version of $\mathbb{P}_C^{Z|YX}$*

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



$$\mathbb{P}_\alpha^{YZ|X} = \quad (138)$$

$$\Rightarrow \mathbb{P}_\alpha^{XYZ} = \triangleleft \mathbb{P}_\alpha^X \quad \begin{array}{c} \text{---} X \\ \text{---} Y \\ \text{---} Z \end{array} \quad \mathbb{P}_\alpha^{YZ|X} \quad (139)$$

$$= \triangleleft \mathbb{P}_\alpha^X \quad \begin{array}{c} \text{---} X \\ \text{---} Y \\ \text{---} Z \end{array} \quad \mathbb{P}_\alpha^{Y|X} \quad \mathbb{P}_\alpha^{Z|(Y|X)} \quad (140)$$

$$= \triangleleft \mathbb{P}_\alpha^{XY} \quad \begin{array}{c} \text{---} X \\ \text{---} Y \\ \text{---} Z \end{array} \quad \mathbb{P}_\alpha^{Z|(Y|X)} \quad (141)$$

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

Theorem 6.6. *Given probability gap model \mathbb{P}_C , X, Y, Z such that $\mathbb{P}_C^{Z|YX}$ exists, $\mathbb{P}_C^{Z|Y}$ exists iff $Z \perp\!\!\!\perp_{\mathbb{P}_C} X|Y$.*

Proof. If: If $Z \perp\!\!\!\perp_{\mathbb{P}_C} X|Y$ then by Theorem 3.23, for each $\mathbb{P}_\alpha \in \mathbb{P}_C$ there exists $\mathbb{P}_\alpha^{Z|Y}$ such that

$$\mathbb{P}_\alpha^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (142)$$

\square

Theorem 6.7 (Valid higher order conditionals). *Suppose we have a sample space (Ω, \mathcal{F}) , variables $X : \Omega \rightarrow X$ and $Y : \Omega \rightarrow Y$, $Z : \Omega \rightarrow Z$ and a probability set \mathbb{P}_C with regular conditional $\mathbb{P}_C^{YZ|X}$, Y discrete and Z standard measurable. Then there exists a valid regular $\mathbb{P}_C^{Z|XY}$.*

Proof. By Theorem 6.4, we have a higher order conditional $\mathbb{P}_C^{Z|(Y|X)}$ which, by Theorem 6.5 is also a version of $\mathbb{P}_C^{Z|XY}$.

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

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

$$\forall A \in \mathcal{Z}, x \in X, y \in Y : \mathbb{P}_C^{YZ|X}(\{y\} \times A|x) = 0 \implies (\mathbb{Q}(A|x, y) = 0) \vee ((X, Y) \bowtie \{(x, y)\} = \emptyset) \quad (143)$$

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

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

6.3 Validity

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

Definition 6.8 (Regular proper conditional probability). Given a probability space $(\mu, \Omega, \mathcal{F})$ and a variable $X : \Omega \rightarrow X$, a regular proper conditional probability $\mu^{|\mathbf{X}} : X \rightarrow \Omega$ is Markov kernel such that

$$\mu(A \cap X^{-1}(B)) = \int_B \mu^{|\mathbf{X}}(A|x) \mu^{\mathbf{X}}(dx) \quad \forall A \in \mathcal{X}, B \in \mathcal{F} \quad (144)$$

$$\iff \quad (145)$$

$$\mu = \triangleleft \mu^{\mathbf{X}} \quad \bullet \quad \boxed{\bar{\mu}^{Y|\mathbf{X}}} \quad \text{---} \quad Y \quad \text{---} \quad X \quad (146)$$

and

$$\mu^{|\mathbf{X}}(X^{-1}(A)|x) = \delta_x(A) \quad (147)$$

Lemma 6.9. Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $X : \Omega \rightarrow X$, $Y : \Omega \rightarrow Y$, if there is a regular proper conditional probability $\mu^{|\mathbf{X}} : X \rightarrow \Omega$ then there is a valid conditional distribution $\mu^{Y|\mathbf{X}}$.

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

Defining $O := \text{id}_{\Omega}$ (the identity function $\Omega \rightarrow \Omega$), $\mu^{|\mathbf{X}}$ is a version of $\mu^{O|\mathbf{X}}$. Note also that $Y = Y \circ O$. Thus by Lemma 6.2, \mathbb{K} is a version of $\mu^{Y|\mathbf{X}}$.

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

$$\mathbb{K}(Y \bowtie A|x) = \delta_x(Y^{-1}(A)) \quad (148)$$

$$= 0 \quad (149)$$

Thus \mathbb{K} is valid. \square

Theorem 6.10 (Validity). *Given (Ω, \mathcal{F}) , $X : \Omega \rightarrow X$, $\mathbb{J} \in \Delta(X)$ with Ω and X standard measurable, there exists some $\mu \in \Delta(\Omega)$ such that $\mu^X = \mathbb{J}$ 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 $X \bowtie x = \emptyset$ but $\mathbb{J}(x) > 0$. Then

$$\mu^X(x) = \mu(X \bowtie x) \quad (150)$$

$$= \sum_{x' \in X} \mathbb{J}(x') \mathbb{K}(X \bowtie x|x') \quad (151)$$

$$= 0 \quad (152)$$

$$\neq \mathbb{J}(x) \quad (153)$$

\square

Lemma 6.11 (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 $\mathbb{P}_C^{Y|X[M]}$ and $\mathbb{Q}_{\{\}}^{Z|YX[M]}$ then defining*

$$\mathbb{R}_{\{\}}^{YZ|X} := \mathbb{P}_C^{Y|X} \odot \mathbb{Q}_{\{\}}^{Z|YX} \quad (154)$$

we have

$$\mathbb{R}_{\{\}}^{YZ|X[M]} = \mathbb{P}_C^{Y|X[M]} \cap \mathbb{Q}_{\{\}}^{Z|YX[M]} \quad (155)$$

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

$$\mathbb{R}_a^{XYZ} = \mathbb{R}_a^X \odot \mathbb{P}_C^{Y|X} \odot \mathbb{Q}_{\{\}}^{Z|YX} \quad (156)$$

$$\implies \mathbb{R}_a^{XY} = \mathbb{R}_a^X \odot \mathbb{P}_C^{Y|X} \quad (157)$$

$$\wedge \mathbb{R}_a^{XYZ} = \mathbb{R}_a^{XY} \odot \mathbb{Q}_{\{\}}^{Z|YX} \quad (158)$$

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

Suppose there's an element \mathbb{S} of $\mathbb{P}_C \cap \mathbb{Q}_{\{\}}$ 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} , $\mathbb{P}_C^{Y|X}$ 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, $\mathbb{P}_C^{Y|X} \odot \mathbb{Q}_{\{\}}^{Z|YX}$ 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 6.12 (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

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

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

$$\forall A \in \mathcal{X} : X \bowtie A = \emptyset \implies \mathbb{P}^X(A) = 0 \quad (160)$$

\square

Lemma 6.13 (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 $C \in \mathcal{Z}$

$$\mathbb{R}^{YZ|X}(B \times C|x) = \int_B \mathbb{Q}^{Z|YX}(C|y, x) \mathbb{P}^{Y|X}(dy|x) \quad (161)$$

$$\leq \mathbb{P}^{Y|X}(B|x) \quad (162)$$

$$= 0 \quad (163)$$

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:

$$\mathbb{R}^{YZ|X}(B \times C|x) = \int_B \mathbb{Q}^{Z|YX}(C|y, x) \mathbb{P}^{Y|X}(dy|x) \quad (164)$$

$$= 0 \quad (165)$$

\square

Corollary 6.14 (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 6.13 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 6.12. \square

Theorem 6.15 (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 6.14.

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

$$(\mathbb{V}^X \odot \mathbb{T}^{Y|X})(x, y) = \mathbb{T}^{Y|X}(y|x) \mathbb{V}^X(x) \quad (166)$$

$$= \mathbb{T}^{Y|X}(y|x) \quad (167)$$

$$> 0 \quad (168)$$

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

6.4 Conditional independence

Theorem ??. *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*

$$\mathbb{P}_C^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (169)$$

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

$$\mathbb{P}_\beta^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (170)$$

And so, by Theorem 3.23, $Y \perp\!\!\!\perp_{\mathbb{P}_\beta} X|W$ for all $\beta \in A$, and in particular $Y \perp\!\!\!\perp_{\mathbb{P}_\alpha} X|W$. Only if: By Theorem 3.23, there exists a version of $\mathbb{P}_\alpha^{Y|WX}$ such that

$$\mathbb{P}_\alpha^{Y|WX} = \begin{array}{c} W \text{ --- } \boxed{\mathbb{P}_\alpha^{Y|W}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (171)$$

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

$$\begin{array}{c} \mathbb{P}_C^{\text{Y|WX}} = \text{W} \text{ --- } \boxed{\mathbb{P}_\alpha^{\text{Y|W}}} \text{ --- } \text{Y} \\ \text{X} \text{ --- } * \end{array} \quad (172)$$

□

This result can fail to hold in the absence of the domination condition. Consider A a collection of inserts that all deterministically set a variable X ; then for any variable Y $Y \perp_{\mathbb{P}_\square} X$ because X is deterministic for any $\alpha \in A$. But $\mathbb{P}_\square^{\text{Y|X}}$ is not necessarily unresponsive to X .

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

6.5 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 $\mathbb{P}^{\text{XY}} \in \Delta(X \times Y)$. Using the standard notion of “hard interventions on X ”, this model induces a probability set which we could informally describe as the set $\mathbb{P}_\square := \{\mathbb{P}_a^{\text{XY}} | a \in X \cup \{*\}\}$ where $*$ is a special element corresponding to the observational setting. The graph $X \longrightarrow Y$ implies the existence of the uniform conditional probability $\mathbb{P}_\square^{\text{Y|X}}$ under the nominated set of interventions, while the usual rules of hard interventions imply that $\mathbb{P}_a^{\text{X}} = \delta_a$ 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 6.16 (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 6.17 (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:

$$\forall B \in \mathcal{Y}, x \in X : (X, Y) \bowtie \{x\} \times B = \emptyset \implies (\mathbb{L}(B|x) = 0) \vee (X \bowtie \{x\} = \emptyset) \quad (173)$$

Definition 6.18 (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 $\mathbb{P}_C^{Y|X[M]}$ associated with \mathbb{L} is the probability set such that for all $\mathbb{P}_\alpha \in \mathbb{P}_C$, \mathbb{L} is a version of $\mathbb{P}_\alpha^{Y|X}$.

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

Lemma 6.13 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 6.15).

Example 6.19. 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 $\mathcal{B} = \frac{\mathcal{W}}{\mathcal{H}^2}$ - i.e. we figure out someone's body mass index first by measuring both their height and weight, and then passing the result through a function that divides the second by the square of the first. Thus, given the random variables W, H modelling \mathcal{W}, \mathcal{H} , \mathcal{B} is the function given by $B = \frac{W}{H^2}$.

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

$$\mathbb{P}_{\square}^{B|WHC} = \begin{array}{c} H \text{ ---} * \\ C \text{ -----} B \\ W \text{ ---} * \end{array} \quad (174)$$

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

$$(B, W, H) \bowtie \{(x, w, h)\} = \{\omega | (W, H)(\omega) = (w, h), B(\omega) = \frac{w}{h^2}\} \quad (175)$$

$$= \emptyset \quad (176)$$

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 174. 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 174 would be used to evaluate the causal effect of body mass index in the causal diagram found in Shahar (2009).

6.6 Causal contractibility

Theorem 4.4. *Exchange commutativity does not imply locality of consequences or vice versa.*

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

Suppose $D = Y = \{0, 1\}$ and we have a probability set \mathbb{P}_C with uniform conditional $\mathbb{P}_C^{Y|D}$, where $D = (D_1, D_2)$, $Y = (Y_1, Y_2)$.

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

$$\mathbb{P}_C^{Y|D}(y_1, y_2 | d_1, d_2) = \mathbb{I}[(y_1, y_2) = (d_1 + d_2, d_1 + d_2)] \quad (177)$$

then

$$\mathbb{P}_C^{Y_1|D}(y_1|d_1, d_2) = \llbracket y_1 = d_1 + d_2 \rrbracket \quad (178)$$

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

$$\text{swap}_{\rho(D)} \mathbb{P}_C^{Y|D}(y_1, y_2|d_1, d_2) = \mathbb{P}_C^{Y|D}(y_1, y_2|d_2, d_1) \quad (179)$$

$$= \llbracket (y_1, y_2) = (d_2 + d_1, d_2 + d_1) \rrbracket \quad (180)$$

$$= \llbracket (y_1, y_2) = (d_1 + d_2, d_1 + d_2) \rrbracket \quad (181)$$

$$= \llbracket (y_2, y_1) = (d_1 + d_2, d_1 + d_2) \rrbracket \quad (182)$$

$$= \mathbb{P}_C^{Y|D} \text{swap}_{\rho(Y)}(y_1, y_2|d_1, d_2) \quad (183)$$

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

$$\mathbb{P}_C^{Y|D}(y_1, y_2|d_1, d_2) = \llbracket (y_1, y_2) = (0, 1) \rrbracket \quad (184)$$

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

$$\mathbb{P}_C^{Y_1|D}(y_1|d_1, d_2) = \llbracket y_1 = 0 \rrbracket \quad (185)$$

$$= \mathbb{P}_C^{Y_1|D_1}(y_1|d_1) \quad (186)$$

$$\mathbb{P}_C^{Y_2|D}(y_2|d_1, d_2) = \llbracket y_2 = 1 \rrbracket \quad (187)$$

$$= \mathbb{P}_C^{Y_2|D_2}(y_2|d_2) \quad (188)$$

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

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

$$\text{swap}_{\rho(D)} \mathbb{P}_C^{Y|D}(y_1, y_2|d_1, d_2) = \mathbb{P}_C^{Y|D}(y_1, y_2|d_2, d_1) \quad (189)$$

$$= \llbracket (y_1, y_2) = (0, 1) \rrbracket \quad (190)$$

$$\neq \llbracket (y_2, y_1) = (0, 1) \rrbracket \quad (191)$$

$$= \mathbb{P}_C^{Y|D} \text{swap}_{\rho(D)}(y_1, y_2|d_1, d_2) \quad (192)$$

□

Theorem 4.7. *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

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \triangle \\ \mu^{Y^D} \\ D \text{ --- } \square \text{F}_{\text{ev}} \text{ --- } Y \end{array} \quad (193)$$

$$\iff \quad (194)$$

$$\mathbb{P}_C^{Y|D}(y|(d_i)_{i \in \mathbb{N}}) = \mu^{Y^D} \Pi_{(d_i i)_{i \in \mathbb{N}}}(y) \quad (195)$$

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

$$\text{ev} : D^{\mathbb{N}} \times Y^{D \times \mathbb{N}} \rightarrow Y \quad (196)$$

$$((d_i)_{i \in \mathbb{N}}, (y_{ij})_{i \in D, j \in \mathbb{N}}) \mapsto (y_{d_i i})_{i \in \mathbb{N}} \quad (197)$$

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

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

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

$$\mathbb{P}_\alpha^{DY}(B \times C) = \int_B \mathbb{P}_C^{Y|D}(C|d) \mathbb{P}_\alpha^D(dd) \quad (198)$$

$$= \int_B \int_D \mathbb{P}_C^{Y|D}(C|d') \mathbb{P}_d^D(dd') \mathbb{P}_\alpha^D(dd) \quad (199)$$

$$= \int_B \mathbb{P}_d^Y(C) \mathbb{P}_\alpha^D(dd) \quad (200)$$

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 $e_{|D|i+j}$ is the i th element of D for all $i, j \in \mathbb{N}$. Define

$$\mu^{Y^D}((y_{ij})_{D \times \mathbb{N}}) := \mathbb{P}_e^Y((y_{|D|i+j})_{i \in D, j \in \mathbb{N}}) \quad (201)$$

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

$$\mathbb{Q} : D \rightarrow Y \quad (202)$$

$$\mathbb{Q} := \begin{array}{c} \triangle \\ \mu^{Y^D} \\ D \text{ --- } \square \text{F}_{\text{ev}} \text{ --- } Y \end{array} \quad (203)$$

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

$$\sum_{y \in Y^{-1}(y_A)} \mathbb{Q}(y|d) = \sum_{y \in Y^{-1}(y_A)} \mu^{(Y_{d_i}^D)^A}(y) \quad (204)$$

$$= \sum_{y \in Y^{-1}(y_A)} \mathbb{P}_e^{(Y|D|d_i+i)^A}(y) \quad (205)$$

$$= \mathbb{P}_e^{\mathbf{Y}_B}(y_A) \quad (206)$$

$$= \mathbb{P}_d^{Y_A}(y_A) \quad \text{by causal contractibility} \quad (207)$$

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

$$\mathbb{Q}(y|d) = \mathbb{P}_d^{\mathbf{Y}}(y) \quad (208)$$

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

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

$$\mu^{\mathbf{Y}^D} \Pi_S = \mathbb{P}_e^{\mathbf{Y}_{\rho(S)}} \quad (209)$$

$$= \mathbb{P}_e^{\mathbf{Y}_{\rho(T)}} \quad \text{by contractibility of } \mathbb{P}_C \text{ and the equality } e_{\rho(S)} = e_{\rho(T)} \quad (210)$$

$$= \mu^{Y^D} \Pi_T \quad (211)$$

If: Suppose

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \text{Diagram showing a triangular block labeled } \mu^{Y^D} \text{ connected to a square block labeled } F_{\text{ev}}. \\ \text{Input } D \text{ enters both blocks from below. The output of } F_{\text{ev}} \text{ is } Y. \end{array} \quad (212)$$

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

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

By definition,

$$\mathbb{P}_C^{\mathbf{Y}_S|D}(y_S|d) = \sum_{y_S^D \in Y^{|D|} \times |S|} \mu^{\mathbf{Y}^D} \Pi_S(y_S^D) \mathbb{F}_{\text{ev}}(y_S|d, y_S^D) \quad (213)$$

$$= \sum_{y_S^D \in Y^{|D| \times |T|}} \mathbb{P}_C^{\mathbf{Y}_T^D}(y_S^D) \mathbb{F}_{\text{ev}}(y_S|d, y_S^D) \quad \text{by contractibility of } \mu^{\mathbf{Y}^D} \Pi_T \quad (214)$$

$$= \mathbb{P}_C^{\mathbf{Y}_T|\mathbf{D}}(y_S|d) \quad (215)$$

1

Lemma 4.8. *Given a $(D; Y)$ -causally contractible model \mathbb{P}'_C on (Ω', \mathcal{F}) , there exists an augmented model \mathbb{P}_C on $(\Omega, \mathcal{F}) := ((\Omega' \times Y^D, \mathcal{F}' \otimes \mathcal{Y}^D))$ such that $\mathbb{P}_C \Pi_{\Omega'} = \mathbb{P}'_C$ and, defining $Y^D : \Omega \times Y^D \rightarrow Y^D$ as the projection onto Y^D*

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \text{triangle containing } \mathbb{P}_C^{Y^D} \\ \text{input } D \text{ from below} \\ \text{output to } \text{box } F_{ev} \\ \text{box } F_{ev} \text{ has input } D \text{ from below and output } Y \end{array} \quad (216)$$

Proof. Set $\mathbb{P}_C^{Y|Y^D D} = \mathbb{F}_{\text{ev}}$ and $\mathbb{P}_C^{Y^D|D} = \mu^{Y^D} \text{erase}_{D^N} \otimes \text{erase}_D$ as in Theorem 4.7. Then Equation 84 follows Theorem 4.7.

Let $W = \Pi_{\Omega'}$ and for each $\alpha \in C$, set

$$\mathbb{P}_\alpha^{\text{W}|Y^D D} = \begin{array}{c} D \\ Y^D \end{array} \rightarrow \boxed{\mathbb{P}_C^{Y|DY^D}} \rightarrow \boxed{\mathbb{P}_\alpha^{\text{W}|DY}} \rightarrow Y \quad (217)$$

Then

$$\mathbb{P}_\alpha^W = \begin{array}{c} \triangleleft \mathbb{P}_\alpha^{D'} \\ \triangleleft \mathbb{P}_C^{Y^D} \end{array} \begin{array}{c} \bullet \\ \text{---} \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y|DY^D}} \\ \text{---} \end{array} \boxed{\mathbb{P}_\alpha^{W|DY}} \text{---} Y \quad (218)$$

$$= \text{Diagram: } \triangleleft \mathbb{P}'^D_\alpha \text{ -- } \bullet \text{ -- } \square \mathbb{P}^Y_C \text{ -- } \square \mathbb{P}'^{W/DY}_\alpha \text{ -- } Y \text{ (with a curved arrow from } \bullet \text{ to } \mathbb{P}'^{W/DY}_\alpha \text{)} \quad (219)$$

$$= \mathbb{P}'_{\alpha} \quad (220)$$

Theorem 4.9. *Suppose we have a sample space (Ω, \mathcal{F}) and a probability set \mathbb{P}_C with uniform conditional $\mathbb{P}_C^{\mathbf{Y}|\mathbf{D}}$ where $\mathbf{D} := (\mathbf{D}_i)_{i \in \mathbb{N}}$ and $\mathbf{Y} := (\mathbf{Y}_i)_{i \in \mathbb{N}}$. \mathbb{P}_C is $(\mathbf{D}; \mathbf{Y})$ -causally contractible if and only if there exists some $\mathbf{H} : \Omega \rightarrow H$ such that $\mathbb{P}_C^{\mathbf{H}}$ and $\mathbb{P}_C^{\mathbf{Y}_i|\mathbf{H}\mathbf{D}_i}$ exist for all $i \in \mathbb{N}$ and*

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \begin{array}{c} \triangle \mu^H \\ \text{---} \bullet \\ \text{---} \bullet \\ \text{---} \end{array} \\ \begin{array}{c} \text{D} \\ \text{---} \end{array} \end{array} \begin{array}{c} \boxed{\Pi_{D,i}} \\ \text{---} \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y_0|HD_0}} \\ \text{---} \end{array} \begin{array}{c} Y_i \\ \text{---} \end{array} \quad (221)$$

$$\rightleftharpoons \quad (222)$$

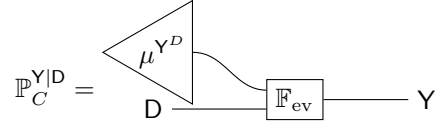
$$Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e Y_{N \setminus i}, D_{N \setminus i} C | H D_i \quad \forall i \in \mathbb{N} \quad (223)$$

$$\wedge H \perp_{\mathbb{P}_C}^e DC \quad (224)$$

$$\wedge \mathbb{P}_G^{\mathbf{Y}_i | \mathbf{HD}_i} = \mathbb{P}^{\mathbf{Y}_0 | \mathbf{HD}_0} \quad \forall i \in \mathbb{N} \quad (225)$$

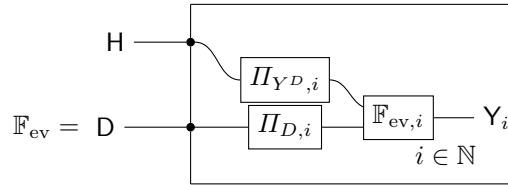
Where $\Pi_{D,i} : D^{\mathbb{N}} \rightarrow D$ is the i th projection map.

Proof. We make use of Lemma 4.7 to show that we can represent the conditional probability $\mathbb{P}_C^{Y|D}$ as



$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \mu^{Y^D} \\ \triangle \\ D \end{array} \xrightarrow{\quad} \boxed{F_{ev}} \xrightarrow{\quad} Y \quad (226)$$

As a preliminary, we will show



$$F_{ev} = D \quad \begin{array}{c} \boxed{\begin{array}{c} H \xrightarrow{\quad} \Pi_{Y^D,i} \xrightarrow{\quad} F_{ev,i} \xrightarrow{\quad} Y_i \\ D \xrightarrow{\quad} \Pi_{D,i} \xrightarrow{\quad} F_{ev,i} \xrightarrow{\quad} Y_i \end{array}} \\ i \in \mathbb{N} \end{array} \quad (227)$$

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 $ev_{Y^D \times D} : Y^D \times D \rightarrow Y$ is the evaluation function

$$((y_i)_{i \in D}, d) \mapsto y_d \quad (228)$$

Recall that ev is the function

$$((d_i)_{i \in \mathbb{N}}, (y_{ij})_{i \in D, j \in \mathbb{N}}) \mapsto (y_{d_i i})_{i \in \mathbb{N}} \quad (229)$$

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

$$F_{ev}(\prod_{i \in \mathbb{N}} A_i | (d_i)_{i \in \mathbb{N}}, (y_{ij})_{i \in D, j \in \mathbb{N}}) = \delta_{(y_{d_i i})_{i \in \mathbb{N}}}(\prod_{i \in \mathbb{N}} A_i) \quad (230)$$

$$= \prod_{i \in \mathbb{N}} \delta_{y_{d_i i}}(A_i) \quad (231)$$

$$= \text{copy}^{\mathbb{N}} \prod_{i \in \mathbb{N}} (\Pi_{D,i} \otimes \Pi_{Y,i}) F_{ev_{Y^D \times D}} \quad (232)$$

Which is what we wanted to show.

Only if: We will assume without loss of generality that we are dealing with an augmented causally contractible model (Lemma 4.8). Because 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 4.7). From [kal \(2005\)](#) we have a di-

recting random measure \mathbf{H} such that

$$\mathbb{P}_C^{\mathbf{Y}^D|\mathbf{H}} = \mathbf{H} \longrightarrow \left[\begin{array}{c} \boxed{\mathbb{P}_{\square}^{\mathbf{Y}^D|\mathbf{H}}} \\ \hline i \in \mathbb{N} \end{array} \right] \mathbf{Y}_i \quad (233)$$

$$\Longleftrightarrow \quad (234)$$

$$\mathbb{P}_C^{\mathbf{Y}^D|\mathbf{H}}\left(\prod_{i \in \mathbb{N}} A_i|h\right) = \prod_{i \in \mathbb{N}} \mathbb{P}_C^{\mathbf{Y}_i^D|\mathbf{H}}(A_i|h) \quad (235)$$

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

$$\mathbb{P}_C^{\mathbf{Y}|\mathbf{H}\mathbf{D}} = \mathbb{P}_C^{\mathbf{Y}^D|\mathbf{H}\mathbf{D}} \odot \mathbb{P}_C^{\mathbf{Y}|\mathbf{Y}^D\mathbf{H}\mathbf{D}} \quad (236)$$

$$\begin{aligned} & \begin{array}{c} \mathbf{H} \longrightarrow \boxed{\mathbb{P}_{\square}^{\mathbf{Y}^D|\mathbf{H}}} \\ \mathbf{D} \longrightarrow \boxed{\mathbb{P}_{\square}^{\mathbf{Y}|\mathbf{Y}^D\mathbf{D}}} \end{array} \longrightarrow \mathbf{Y} \\ &= \quad \quad \quad = \quad \begin{array}{c} \triangleleft \mu^{\mathbf{H}} \triangleleft \\ \mathbf{D} \longrightarrow \left[\begin{array}{c} \boxed{\Pi_{D,i}} \quad \boxed{\mathbb{P}_C^{\mathbf{Y}_i^D|\mathbf{H}\mathbf{D}_0}} \\ \hline i \in \mathbb{N} \end{array} \right] \mathbf{Y}_i \end{array} \end{aligned} \quad (237)$$

If: By assumption

$$\mathbb{P}_C^{\mathbf{Y}|\mathbf{D}}\left(\prod_{i \in \mathbb{N}} A_i|h, (d_i)_{i \in \mathbb{N}}\right) = \int_H \prod_{i \in \mathbb{N}} \mathbb{P}_C^{\mathbf{Y}_i|\mathbf{H}\mathbf{D}_1}(A_i|h, d_i) \mathbb{P}_C^{\mathbf{H}}(dh) \quad (238)$$

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

$$\mathbb{P}_{\alpha}^{\mathbf{Y}_M}(A) = \int_{D^{\mathbb{N}}} \mathbb{P}_{\alpha}^{\mathbf{Y}_M|\mathbf{D}}(A|d) \mathbb{P}_{\alpha}^{\mathbf{D}}(dd) \quad (239)$$

$$= \int_H \int_{D^{\mathbb{N}}} \prod_{i \in M} \mathbb{P}_C^{\mathbf{Y}_i|\mathbf{H}\mathbf{D}_1}(A_i|h, d_i) \mathbb{P}_{\alpha}^{\mathbf{D}}(dd) \mathbb{P}_C^{\mathbf{H}}(dh) \quad (240)$$

$$= \int_H \int_{D^{|M|}} \prod_{i \in M} \mathbb{P}_C^{\mathbf{Y}_i|\mathbf{H}\mathbf{D}_1}(A_i|h, d_i) \mathbb{P}_{\alpha}^{\mathbf{D}^M}(dd_M) \mathbb{P}_C^{\mathbf{H}}(dh) \quad (241)$$

$$= \int_H \int_{D^{|M|}} \prod_{i \in M} \mathbb{P}_C^{\mathbf{Y}_i|\mathbf{H}\mathbf{D}_1}(A_i|h, d_i) \mathbb{P}_{\alpha'}^{\mathbf{D}^N}(dd_N) \mathbb{P}_C^{\mathbf{H}}(dh) \quad (242)$$

$$= \int_H \int_{D^{\mathbb{N}}} \prod_{i \in M} \mathbb{P}_C^{\mathbf{Y}_i|\mathbf{H}\mathbf{D}_1}(A_i|h, d_i) \mathbb{P}_{\alpha'}^{\mathbf{D}}(dd) \mathbb{P}_C^{\mathbf{H}}(dh) \quad (243)$$

$$= \mathbb{P}_{\alpha'}^{\mathbf{Y}_M}(A) \quad (244)$$

□

6.7 Body mass index revisited

Lemma 6.20. *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

$$Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{\{i\}^c}, X_{\{i\}^c}, D_{\{i\}^c}) C | H D_i X_i \quad (245)$$

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

$$Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{\{i\}^c}, X_{\{i\}^c}) C | H X_i \quad (246)$$

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

$$\mathbb{P}_C^{Y_i | X_i D_i H} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_j | X_j D_j H} \quad (247)$$

$$\implies \mathbb{P}_C^{Y_i | X_i H} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_j | X_j H} \quad (248)$$

□

Theorem ??. *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 ?? 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

$$(X_i, Y_i) \perp\!\!\!\perp_{\mathbb{P}_\alpha} (X_{\{i\}^c}, Y_{\{i\}^c}, D_{\{i\}^c}) | H D_i \quad (249)$$

$$Y_i \perp\!\!\!\perp_{\mathbb{P}_\alpha} (Y_{\{i\}^c}, X_{\{i\}^c}) | H D_i X_i \quad (250)$$

Where Eq. 250 follows from 249 by weak union. By Theorem ??, $Y_i \perp\!\!\!\perp_{\mathbb{P}_\alpha} (Y_{\{i\}^c}, X_{\{i\}^c}) | H D_i X_i$ 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 ??. 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 6.20. □

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 Duflo for her leadership on the handbook and for extensive comments on earlier drafts. Chassang and Snowberg gratefully acknowledge the support of NSF grant SES-1156154. In Abhijit Vinayak Banerjee and Esther Duflo, editors, *Handbook of Economic Field Experiments*, volume 1 of *Handbook of Field Experiments*, pages 141–174. North-Holland, January 2017. doi: 10.1016/bs.hefe.2016.08.005. URL <https://www.sciencedirect.com/science/article/pii/S2214658X16300071>.
- Abhijit V. Banerjee, James Berry, Esther Duflo, Harini Kannan, and Shobhini Mukerji. Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of 'Teaching at the Right Level' in India. SSRN Scholarly Paper ID 2843569, Social Science Research Network, Rochester, NY, September 2016. URL <https://papers.ssrn.com/abstract=2843569>.
- Vladimir Bogachev and Ilya Malofeev. Kantorovich problems and conditional measures depending on a parameter. *Journal of Mathematical Analysis and Applications*, 486:123883, June 2020. doi: 10.1016/j.jmaa.2020.123883.
- Stephan Bongers, Jonas Peters, Bernhard Schölkopf, and Joris M. Mooij. Theoretical Aspects of Cyclic Structural Causal Models. *arXiv:1611.06221 [cs, stat]*, November 2016. URL <http://arxiv.org/abs/1611.06221>. arXiv: 1611.06221.
- 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.
- 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.
- Brendan Fong. Causal Theories: A Categorical Perspective on Bayesian Networks. *arXiv:1301.6201 [math]*, January 2013. URL <http://arxiv.org/abs/1301.6201>. arXiv: 1301.6201.
- Tobias Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, August 2020. ISSN 0001-8708. doi: 10.1016/j.aim.2020.107239. URL <https://www.sciencedirect.com/science/article/pii/S0001870820302656>.
- Sander Greenland and James M Robins. Identifiability, Exchangeability, and Epidemiological Confounding. *International Journal of Epidemiology*, 15(3): 413–419, September 1986. ISSN 0300-5771. doi: 10.1093/ije/15.3.413. URL <https://doi.org/10.1093/ije/15.3.413>.
- D. Heckerman and R. Shachter. Decision-Theoretic Foundations for Causal Reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, December 1995. ISSN 1076-9757. doi: 10.1613/jair.202. URL <https://www.jair.org/index.php/jair/article/view/10151>.
- M. A. Hernán and S. L. Taubman. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity*, 32(S3):S8–S14, August 2008. ISSN 1476-5497. doi: 10.1038/ijo.2008.82. URL <https://www.nature.com/articles/ijo200882>.
- Miguel A. Hernán. Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*, 26(10):674–680, October 2016. ISSN 1047-2797. doi: 10.1016/j.annepidem.2016.08.016. URL <http://www.sciencedirect.com/science/article/pii/S1047279716302800>. Publisher: Elsevier.

- 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.
- 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.
- Eyal Shahar. The association of body mass index with health outcomes: causal, inconsistent, or confounded? *American Journal of Epidemiology*, 170(8):957–958, October 2009. ISSN 1476-6256. doi: 10.1093/aje/kwp292.
- Peter 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=tvp&paperid=4710&option_lang=eng.

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: