Causal Statistical Decision Theory|What are interventions?

David Johnston

May 24, 2022

Contents

Chapter 1

Introduction

1.1 Theories of causal inference

Beginning in the 1930s, a number of associations between cigarette smoking and lung cancer were established: on a population level, lung cancer rates rose rapidly alongside the prevalence of cigarette smoking. Lung cancer patients were far more likely to have a smoking history than demographically similar individuals without cancer and smokers were around 40 times as likely as demographically similar non-smokers to go on to develop lung cancer. In laborotory experiments, cells which were introduced to tobacco smoke developed *ciliastasis*, and mice exposed to cigarette smoke tars developed tumors(Proctor, 2012). Nevertheless, until the late 1950s, substantial controversy persisted over the question of whether the available data was sufficient to establish that smoking cigarettes *caused* lung cancer. Cigarette manufacturers famously argued against any possible connection (Oreskes and Conway, 2011) and Roland Fisher in particular argued that the available data was not enough to establish that smoking actually caused lung cancer (Fisher, 1958). Today, it is widely accepted that cigarettes do cause lung cancer, along with other serious conditions such as vascular disease and chronic respiratory disease (World Health Organisation, 2018; Wiblin, 2016).

The question of a causal link between smoking and cancer is a very important one to many different people. Individuals who enjoy smoking (or think they might) may wish to avoid smoking if cigarettes pose a severe health risk, so they are interested in knowing whether or not it is so. Additionally, some may desire reassurance that their habit is not too risky, whether or not this is true. Potential and actual investors in cigarette manufacturers may see health concerns as a barrier to adoption, and also may personally want to avoid supporting products that harm many people. Like smokers, such people might have some interest in knowing the truth of this question, and a separate interest in hearing that cigarettes are not too risky, whether or not this is true. Governments and organisations with a responsibility for public health may see themselves as having responsibility to discourage smoking as much as possible if smoking is severely detrimental to health. The costs and benefits of poor decisions about smoking are large: 8 million annual deaths are attributed to cigarette-caused cancer and vascular disease in 2018(World Health Organisation, 2018) while global cigarette sales were estimated at US$711 billion in 2020 (Statista, 2020) (a figure which might be substantially larger if cigarettes were not widely believed to be harmful).

The question of whether or not cigarette smoking causes cancer illustrates two key facts about causal questions: First, having the right answers to causal questions is of tremendous importance to huge numbers of people. Second, confusion over causal questions can persist even when a great deal of data and facts relevant to the question are agreed upon.

Causal conclusions are often justified on the basis of ad-hoc reasoning. For example Krittanawong et al. (2020) state:

> [...] the potential benefit of increased chocolate consumption, reducing coronary artery disease (CAD) risk is not known. We aimed to explore the association between chocolate consumption and CAD.

It is not clear whether Krittanawong et. al. mean that a negative association between chocolate consumption and CAD implies that increased chocolate consumption is likely to reduce coronary artery disease (which is suggested by the word "benefit"), or that an association may be relevant to the question and the reader should draw their own conclusions. Whether the implication is being suggested by Krittanawong et. al. or merely imputed by naïve readers, it is being drawn on an ad-hoc basis – no argument for the implication can be found in this paper. As Pearl (2009) has forcefully argued, additional assumptions are always required to answer causal questions from associational facts, and stating these assumptions explicitly allows those assumptions to be productively scrutinised.

For causal questions that are controversial or difficult, it is tremendously advantageous to be able to address them transparently. Theories of causation enable this; given a theory of causation and a set of assumptions, if anyone claims that some conclusion follows it is publicly verifiable whether or not it actually does so. If the deduction is correct, then any remaining disagreement must be in the assumptions or in the theory. For people who are interested in understanding what is true, pinpointing disagreement can be enlightening. Someone could learn, for example, that there are assumptions that they find plausible that permit conclusions they did not initially believe. Alternatively, if a motivated conclusion follows only from implausible assumptions, hearing these assumptions explicitly might make the conclusion less attractive.

Theories of causation help us to answer causal questions, which means that before we have any theory, we already have causal questions we want to answer. If potential outcomes notation and causal graphical models had never been invented there would still be just as many people who want to the answer to questions something like "does smoking causes cancer?", even if on-one could say what exactly they meant by "causes" and even if many people actually want answers to slightly different questions. Theories exist to serve our need for transparent answers to causal questions.

Potential outcomes and causal graphical models are prominent examples of "practical theories" of causation. I call them "practical theories" because most of the time we encounter them they are being used to answer "practical" questions like "Does smoking cause cancer?", or "In general, when does data allow us to conclude that X causes Y?" It is less common to see the "fundamental questions" addressed, like "Does the theory of causal graphical models offer an adequate account of what 'cause' means?", which is more often found in the field of philosophy. Spirtes et al. (2000) explain their motivation to study what I call "practical theories of causation" as follows:

> One approach to clarifying the notion of causation – the philosophers approach ever since Plato – is to try to define "causation" in other terms, to provide necessary and sufficient and noncircular conditions for one thing, or feature or event or circumstance, to cause another, the way one can define "bachelor" as "unmarried adult male human." Another approach to the same problem – the mathematicians approach ever since Euclid – is to provide axioms that use the notion of causation without defining it, and to investigate the necessary consequences of those assumptions. We have few fruitful examples of the first sort of clarification, but many of the second [...]

I think what Spirtes, Glymour and Scheines (henceforth: SGS) mean here is that they *define* a notion of causation – because causal graphical models do define a notion of causation – without interrogating whether it means the same thing as the word "causation". Incidentally, since publication of this paragraph, the notion of causation defined by causal graphical models has been subject to substantial interrogation by philosophers (Woodward, 2016).

I am sympathetic to the argument that it does not matter a great deal whether "causal-graphical-models-causation" and "causation" mean the same thing in everyday language. It is common for words to have somewhat different meanings when used by specialists to when they are used by laypeople, and this isn't because the specialists are ignorant or confused about their subject. However, I think it matters a lot which causal questions can be transparently answered by "causal-graphical-models-causation", and so I believe that the notions of causation adopted by practical theories do warrant scrutiny.

I think one reason that SGS are keen to avoid dwelling on the definition of causation is that satisfactory definitions of causation are difficult. For example, causal graphical models depend on the notion of *causal relationships* between variables. These may be defined as follows:

> X_i is a *cause* of X_j if there is an *ideal intervention* on X_i that changes the value X_j

This definition is incomplete without a definition of "ideal interventions". Ideal interventions may be defined by their action in "causally sufficient models":

- An $[\mathsf{X}_i, \mathsf{X}_j]$-ideal intervention is an operation whose result is determined by applying the *do-calculus* to a *causally sufficient* model $((\Omega, \mathcal{F}, \mathbb{P}), \mathcal{G}, \boldsymbol{U})$
- A model $((\Omega, \mathcal{F}, \mathbb{P}), \mathcal{G}, \boldsymbol{U})$ is $[\mathsf{X}_i, \mathsf{X}_j]$-causally sufficient if U contains X_i, X_j and "all intervenable variables that *cause*" both X_i and X_j [1]

While I don't offer a definition of the *do-calculus* in this introduction, it can be rigorously defined, see for example Pearl (2009). The problem is that the definition of a *causally sufficient* model itself invokes the word *cause*, which is what the original definition was trying to address. Circularity is a recognised problem with interventional definitions of causation (Woodward, 2016). In Section **??**, I further show models with ideal interventions generally have counterintuitive properties. The purpose of a theory of causation like causal graphical models is to support transparent reasoning about causal questions, and a circular definition that leads to counterintuitive conclusions undermines this purpose.

As with Euclid's parallel postulate, I think it is reasonable to ask if the notion of ideal interventions and other causal definitions can be modified or avoided. Causal statistical decision theory (CSDT) is a theory of causation that is motivated by the problem of *what is generally needed to answer causal questions* rather than *what does "causation" mean?* Along similar lines to CSDT, Dawid (2020) has observed that the problem of deciding how to act in light of data can be formalised without appeal to theories of causation. We develop this in substantial detail, showing how both *interventional models* and *counterfactual models* arise as special cases of CSDT.

I want to revisit the claims about what I actually show, hopefully to add to it

A key feature of CSDT is what I call the *option set*. This is the set of decisions, acts or counterfactual propositions under consideration in a given problem. A causal graphical model and a potential outcomes model will both implicitly define an option set as a result of their basic definitions of causation, but CSDT demands that this is done explicitly. I argue that this is a key strength of CSDT, on the basis of the following claims which I defend in the following chapters:

- Causal questions are not well-posed without an option set in the same way a function is not well-defined without its domain
- The option set need not correspond in any fixed manner to the set of observed variables
- The nature of the option set can affect the difficulty of causal inference questions

I commented out an additional section about potential outcomes and closest world counterfactuals, which is a second example of "opaque causal definitions". I'm interested if any readers think it would be good to have a second example

[1] Weaker conditions for causal sufficiency are possible, but they don't avoid circularity (Shpitser and Pearl, 2008)

Chapter 2

Technical Prerequisites

Our approach to causal inference is (like most other approaches) based on probability theory. Many results and conventions will be familiar to readers, and these are collected in Section 2.2.1.

Less likely to be familiar to readers is the string diagram notation we use to represent probabilistic functions. This is a notation 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. It is introduced in Section 2.3.

We also extend the theory of probability to a theory of probability sets, which we introduce in Section 2.4. This section goes over some ground already trodden by Section 2.2.1; this structure was chosen so that people familiar with the Section 2.2.1 can skip to Section 2.4 for relevant generalisations to probability sets. Two key ideas introduced here are *uniform conditional probability*, similar but not identical to conditional probability, and *extended conditional independence* as introduced by Constantinou and Dawid (2017), similar but not identical to regular conditional independence.

We finally introduce the assumption of *validity*, which ensures that probability sets constructed by "assembling" collections of uniform conditionals are non-empty.

This is a reference chapter – a reader who is already quite familiar with probability theory may skip to Chapter 3. Where necessary, references back to theorems and definitions in this chapter are given. In Chapter 4, we will introduce one additional probabilistic primitive: *combs*, as we feel that additional context is helpful for understanding them.

2.1 Conventions

One of the unusual conventions in this thesis is the notation of uniform conditional probability. Given a set of probability distributions $\mathbb{P}_C := \{\mathbb{P}_\alpha | \alpha \in C\}$ on a common sample space $(\Omega, \mathcal{F})$ with variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ represents a Markov kernel $X \to Y$ that satisfies the definition of the distribution of Y given X (Definition 2.2.16) for every $\alpha \in C$, while $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}$ is a conditional distribution only for α. There are two unusual feature: firstly, it is more common to write a conditional distribution $\mathbb{P}(\mathsf{Y}|\mathsf{X})$ and secondly, the subscript indicating the "domain of validity" of the conditional probability is unusual.

Because this thesis uses sets of probability measures rather than single probability measures, in general a conditional distribution may be valid only for some subset of the probability measures, and always including a subscript indicating which subset or element for which a conditional distribution is valid avoids any ambiguity about this. Avoiding notation of the form $\mathbb{P}(\mathsf{Y}|\mathsf{X})$ is an aesthetic preference; writing a conditional distribution like this suggests $\mathbb{P}(\mathsf{Y}|\mathsf{X})$ is the result of function composition between $\mathbb{P}$ and some function denoted "$\mathsf{Y}|\mathsf{X}$". However, conditional probabilities are not given by composition of functions like this.

Name	notation	meaning
Iverson bracket	$[\![\cdot]\!]$	Function equal to 1 if $\cdot$ is true, false otherwise
Identity function	idf_X	Identity function $X \to X$
Identity kernel	id_X	Kernel associated with the identity function $X \to X$

2.2 Probability Theory

2.2.1 Standard Probability Theory

σ-algebras

Definition 2.2.1 (Sigma algebra)**.** Given a set A, a σ-algebra $\mathcal{A}$ is a collection of subsets of A where

- $A \in \mathcal{A}$ and $\emptyset \in \mathcal{A}$
- $B \in \mathcal{A} \implies B^C \in \mathcal{A}$
- $\mathcal{A}$ is closed under countable unions: For any countable collection $\{B_i | i \in Z \subset \mathbb{N}\}$ of elements of $\mathcal{A}$, $\cup_{i \in Z} B_i \in \mathcal{A}$

Definition 2.2.2 (Measurable space)**.** A measurable space $(A, \mathcal{A})$ is a set A along with a σ-algebra $\mathcal{A}$.

Definition 2.2.3 (Sigma algebra generated by a set)**.** Given a set A and an arbitrary collection of subsets $U \supset \mathcal{P}(A)$, the σ-algebra generated by U, $\sigma(U)$, is the smallest σ-algebra containing U.

Common σ algebras For any A, $\{\emptyset, A\}$ is a σ-algebra. In particular, it is the only sigma algebra for any one element set $\{*\}$.

For countable A, the power set $\mathcal{P}(A)$ is known as the discrete σ-algebra.

Given A and a collection of subsets of $B \subset \mathcal{P}(A)$, $\sigma(B)$ is the smallest σ-algebra containing all the elements of B.

If A is a topological space with open sets T, $\mathcal{B}(\mathbb{R}) := \sigma(T)$ is the *Borel σ-algebra* on A.

If A is a separable, completely metrizable topological space, then $(A, \mathcal{B}(A))$ is a *standard measurable set*. All standard measurable sets are isomorphic to either $(\mathbb{R}, B(\mathbb{R}))$ or $(C, \mathcal{P}(C))$ for denumerable C (Çinlar, 2011, Chap. 1).

Probability measures and Markov kernels

Definition 2.2.4 (Probability measure)**.** Given a measurable space $(E, \mathcal{E})$, a map $\mu : \mathcal{E} \to [0, 1]$ is a *probability measure* if

- $\mu(E) = 1$, $\mu(\emptyset) = 0$
- Given countable collection $\{A_i\} \subset \mathcal{E}$, $\mu(\cup_i A_i) = \sum_i \mu(A_i)$

Nxample 2.2.5 (Set of all probability measures)**.** The set of all probability measures on $(E, \mathcal{E})$ is written $\Delta(E)$.

Definition 2.2.6 (Markov kernel)**.** Given measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$, a *Markov kernel* or *stochastic function* is a map $\mathbb{M} : E \times \mathcal{F} \to [0, 1]$ such that

- The map $\mathbb{M}(A|\cdot) : x \mapsto \mathbb{M}(A|x)$ is $\mathcal{E}$-measurable for all $A \in \mathcal{F}$
- The map $\mathbb{M}(\cdot|x) : A \mapsto \mathbb{M}(A|x)$ is a probability measure on $(F, \mathcal{F})$ for all $x \in E$

Nxample 2.2.7 (Signature of a Markov kernel)**.** Given measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$ and $\mathbb{M} : E \times \mathcal{F} \to [0, 1]$, we write the signature of $\mathbb{M} : E \rightarrowtriangle F$, read "$\mathbb{M}$ maps from E to probability measures on F".

Definition 2.2.8 (Deterministic Markov kernel)**.** A *deterministic* Markov kernel $\mathbb{A} : E \to \Delta(\mathcal{F})$ is a kernel such that $\mathbb{A}_x(B) \in \{0, 1\}$ for all $x \in E$, $B \in \mathcal{F}$.

Common probability measures and Markov kernels

Definition 2.2.9 (Dirac measure)**.** The *Dirac measure* $\delta_x \in \Delta(X)$ is a probability measure such that $\delta_x(A) = [\![x \in A]\!]$

Definition 2.2.10 (Markov kernel associated with a function)**.** Given measurable $f : (X, \mathcal{X}) \to (Y, \mathcal{Y})$, $\mathbb{F}_f : X \rightarrowtriangle Y$ is the Markov kernel given by $x \mapsto \delta_{f(x)}$

Definition 2.2.11 (Markov kernel associated with a probability measure)**.** Given $(X, \mathcal{X})$, a one-element measurable space $(\{*\}, \{\{*\}, \emptyset\})$ and a probability measure $\mu \in \Delta(X)$, the associated Markov kernel $\mathbb{Q}_\mu : \{*\} \rightarrowtriangle X$ is the unique Markov kernel $* \mapsto \mu$

Lemma 2.2.12 (Products of functional kernels yield function composition)**.** *Given measurable $f : X \to Y$ and $g : Y \to Z$, $\mathbb{F}_f \mathbb{F}_g = \mathbb{F}_{g \circ f}$.*

Proof.

$$(\mathbb{F}_f \mathbb{F}_g)_x(A) = \int_X (\mathbb{F}_g)_y(A) d(\mathbb{F}_f)_x(y) \tag{2.1}$$

$$= \int_X \delta_{g(y)}(A) d\delta_{f(x)}(y) \tag{2.2}$$

$$= \delta_{g(f(x))}(A) \tag{2.3}$$

$$= (\mathbb{F}_{g \circ f})_x(A) \tag{2.4}$$

□

Variables, conditionals and marginals

Definition 2.2.13 (Variable)**.** Given a measurable space $(\Omega, \mathcal{F})$ and a measurable space of values $(X, \mathcal{X})$, an *X-valued variable* is a measurable function $\mathsf{X} : (\Omega, \mathcal{F}) \to (X, \mathcal{X})$.

Definition 2.2.14 (Sequence of variables)**.** Given a measurable space $(\Omega, \mathcal{F})$ and two variables $\mathsf{X} : (\Omega, \mathcal{F}) \to (X, \mathcal{X})$, $\mathsf{Y} : (\Omega, \mathcal{F}) \to (Y, \mathcal{Y})$, $(\mathsf{X}, \mathsf{Y}) : \Omega \to X \times Y$ is the variable $\omega \mapsto (\mathsf{X}(\omega), \mathsf{Y}(\omega))$.

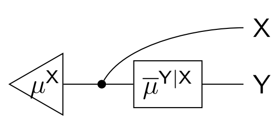
Definition 2.2.15 (Marginal distribution)**.** Given a probability space $(\mu, \Omega, \mathcal{F})$ and a variable $\mathsf{X} : \Omega \to (X, \mathcal{X})$, the *marginal distribution* of X with respect to μ, $\mu^{\mathsf{X}} : \mathcal{X} \to [0, 1]$ by $\mu^{\mathsf{X}}(A) := \mu(\mathsf{X}^{-1}(A))$ for any $A \in \mathcal{X}$.

Definition 2.2.16 (Conditional distribution)**.** Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, the *conditional distribution* of Y given X is any Markov kernel $\mu^{\mathsf{Y}|\mathsf{X}} : X \to Y$ such that

$$\mu^{\mathsf{XY}}(A \times B) = \int_A \mu^{\mathsf{Y}|\mathsf{X}}(B|x) \mathrm{d}\mu^{\mathsf{X}}(x) \qquad \forall A \in \mathcal{X}, B \in \mathcal{Y} \tag{2.5}$$

$$\iff \tag{2.6}$$

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



(2.7)

Markov kernel product notation

Three pairwise *product* operations involving Markov kernels can be defined: measure-kernel products, kernel-kernel products and kernel-function products. These are analagous to row vector-matrix products, matrix-matrix products and matrix-column vector products respectively.

, $\mathsf{T} : Y \to T$, $\mathbb{M} : X \to \Delta(\mathcal{Y})$ and $\mathbb{N} : Y \to \Delta(\mathcal{Z})$

Definition 2.2.17 (Measure-kernel product)**.** Given $\mu \in \Delta(\mathcal{X})$ and $\mathbb{M} : X \rightarrowtriangle Y$, the *measure-kernel product* $\mu\mathbb{M} \in \Delta(Y)$ is given by

$$\mu\mathbb{M}(A) := \int_X \mathbb{M}(A|x)\mu(\mathrm{d}x) \tag{2.8}$$

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

Definition 2.2.18 (Kernel-kernel product)**.** Given $\mathbb{M} : X \rightarrowtriangle Y$ and $\mathbb{N} : Y \rightarrowtriangle Z$, the *kernel-kernel product* $\mathbb{M}\mathbb{N} : X \rightarrowtriangle Z$ is given by

$$\mathbb{M}\mathbb{N}(A|x) := \int_Y \mathbb{N}(A|x)\mathbb{M}(\mathrm{d}y|x) \tag{2.9}$$

for all $A \in \mathcal{Z}$, $x \in X$.

Definition 2.2.19 (Kernel-function product)**.** Given $\mathbb{M} : X \rightarrowtriangle Y$ and $f : Y \to Z$, the *kernel-function product* $\mathbb{M}f : X \to Z$ is given by

$$\mathbb{M}f(x) := \int_Y f(y)\mathbb{N}(\mathrm{d}y|x) \tag{2.10}$$

for all $x \in X$.

Definition 2.2.20 (Tensor product)**.** Given $\mathbb{M} : X \rightarrowtriangle Y$ and $\mathbb{L} : W \rightarrowtriangle Z$, the tensor product $\mathbb{M} \otimes \mathbb{N} : X \times W \rightarrowtriangle Y \times Z$ is given by

$$(\mathbb{M} \otimes \mathbb{L})(A \times B|x, w) := \mathbb{M}(A|x)\mathbb{L}(B|w) \tag{2.11}$$

For all $x \in X$, $w \in W$, $A \in \mathcal{Y}$ and $B \in \mathcal{Z}$.

All products are associative (Çinlar, 2011, Chapter 1).

One application of the product notation is that marginal distributions can be alternatively defined in terms of a kernel product, as shown in Lemma 2.2.21.

Lemma 2.2.21 (Marginal distribution as a kernel product)**.** *Given a probability space* $(\mu, \Omega, \mathcal{F})$ *and a variable* $\mathsf{X} : \Omega \to (X, \mathcal{X})$*, define* $\mathbb{F}_\mathsf{X} : \Omega \rightarrowtriangle X$ *by* $\mathbb{F}_\mathsf{X}(A|\omega) = \delta_{\mathsf{X}(\omega)}(A)$*, then*

$$\mu^\mathsf{X} = \mu\mathbb{F}_\mathsf{X} \tag{2.12}$$

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

$$\mu\mathbb{F}_\mathsf{X}(A) = \int_\Omega \delta_{\mathsf{X}(\omega)}(A)\mathrm{d}\mu(\omega) \tag{2.13}$$

$$= \int_{\mathsf{X}^{-1}(\omega)} \mathrm{d}\mu(\omega) \tag{2.14}$$

$$= \mu^\mathsf{X}(A) \tag{2.15}$$

□

Semidirect product

Given a marginal μ^{X} and a conditional $\mu^{\mathsf{Y}|\mathsf{X}}$, the product of the two yields the marginal distribution of Y: $\mu^{\mathsf{Y}} = \mu^{\mathsf{X}}\mu^{\mathsf{Y}|\mathsf{X}}$. We define another product – the *semidirect* product $\odot$ – as the product that yields the joint distribution of (X, Y): $\mu^{\mathsf{XY}} = \mu^{\mathsf{X}} \odot \mu^{\mathsf{Y}|\mathsf{X}}$. The semidirect product is associative (Lemma 2.2.23)

Definition 2.2.22 (Semidirect product)**.** Given $\mathbb{K} : X \rightarrowtriangle Y$ and $\mathbb{L} : Y \times X \rightarrowtriangle Z$, the semidirect product $\mathbb{K} \odot \mathbb{L} : X \to Y \times Z$ is given by

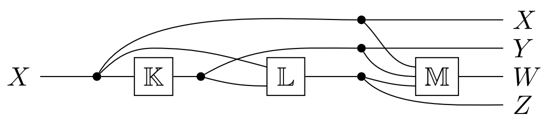
$$(\mathbb{K} \odot \mathbb{L})(A \times B|x) = \int_A \mathbb{L}(B|y, x)\mathbb{K}(dy|x) \qquad \forall A \in \mathcal{Y}, B \in \mathcal{Z} \tag{2.16}$$

Lemma 2.2.23 (Semidirect product is associative)**.** *Given* $\mathbb{K} : X \rightarrowtriangle Y$, $\mathbb{L} : Y \times X \rightarrowtriangle Z$ *and* $\mathbb{M} : Z \times Y \times X \rightarrowtriangle W$

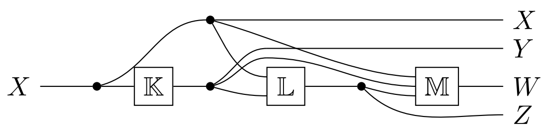
$$(\mathbb{K} \odot \mathbb{L}) \odot \mathbb{Z} = \mathbb{K} \odot (\mathbb{L} \odot \mathbb{Z}) \tag{2.17}$$
$$\tag{2.18}$$

Proof.

$$(\mathbb{K} \odot \mathbb{L}) \odot \mathbb{M} = \tag{2.19}$$



$$= \tag{2.20}$$



$$= \mathbb{K} \odot (\mathbb{L} \odot \mathbb{M}) \tag{2.21}$$

□

The semidirect product can be used to define a notion of almost sure equality: two kernels $\mathbb{K} : X \rightarrowtriangle Y$ and $\mathbb{L} : X \rightarrowtriangle Y$ are μ-almost surely equal if $\mu \odot \mathbb{K} = \mu \odot \mathbb{L}$. This is identical to the notion of almost sure equality in Cho and Jacobs (2019), who shows that under the assumption that $(Y, \mathcal{Y})$ is countably generated, $\mathbb{K} \overset{\mu}{\cong} \mathbb{L}$ if and only if $\mathbb{K} = \mathbb{L}$ μ-almost everywhere.

Definition 2.2.24 (Almost sure equality)**.** Two Markov kernels $\mathbb{K} : X \rightarrowtriangle Y$ and $\mathbb{L} : X \rightarrowtriangle Y$ are almost surely equal $\overset{\mathbb{P}_C}{\cong}$ with respect to a probability space $(\mu, X, \mathcal{X})$, written $\mathbb{K} \overset{\mu}{\cong} \mathbb{L}$ if

$$\mu \odot \mathbb{K} = \mu \odot \mathbb{L} \tag{2.22}$$

Theorem 2.2.25. *Given* $(\mu, X, \mathcal{X})$, $\mathbb{K} : X \rightarrowtriangle Y$ *and* $\mathbb{L} : X \rightarrowtriangle Y$, $\mathbb{K} \overset{\mu}{\cong} \mathbb{L}$ *if and only if, defining* $U := \{x | \exists A \in \mathcal{Y} : \mathbb{K}(A|x) \neq \mathbb{L}(A|x)\}$, $\mu(U) = 0$.

Proof. Cho and Jacobs (2019) proposition 5.4. □

We often want to talk about almost sure equality of two different versions $\mathbb{K}$ and $\mathbb{L}$ of a conditional distribution $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ with respect to some ambient probability space $(\mathbb{P}, \Omega, \mathcal{F})$. This simply means $\mathbb{K}$ and $\mathbb{L}$ satisfy Definition 2.2.16 with respect to $\mathbb{P}$, X and Y, and they are almost surely equal with respect to the marginal $\mathbb{P}^{\mathsf{X}}$. Tthe relevant variables are usually obvious from the context and we leave them implicit and we will write $\mathbb{K} \overset{\mathbb{P}}{\cong} \mathbb{L}$. If the relevant marginal is ambiguous, we will instead write $\mathbb{K} \overset{\mathbb{P}^{\mathsf{X}}}{\cong} \mathbb{L}$

Definition 2.2.26 (Almost sure equality with respect to a pair of variables)**.** Given $(\mathbb{P}, \Omega, \mathcal{F})$ and $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, two Markov kernels $\mathbb{K} : X \rightarrowtriangle Y$ and $\mathbb{L} : X \rightarrowtriangle Y$ are X-almost surely equal with respect to $\mathbb{P}$, written $\mathbb{K} \overset{\mathbb{P}}{\cong} \mathbb{L}$, if they are almost surely equal with respect to the marginal $\mathbb{P}^{\mathsf{X}}$.

2.3 String Diagrams

We make use of 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 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).

2.3.1 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:

$$\mathbb{K} := \ -\boxed{\mathbb{K}}- \tag{2.23}$$

$$\mu := \lhd\!\mathbb{P}\!- \tag{2.24}$$

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

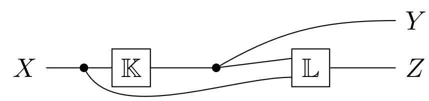
$$\mathbb{L}\mathbb{M} := X\ \boxed{\mathbb{K}}\!-\!\boxed{\mathbb{M}}\ Z \tag{2.25}$$

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

$$\mathbb{K} \otimes \mathbb{L} := \begin{array}{l} W\ \boxed{\mathbb{K}}\ Y \\ X\ \boxed{\mathbb{L}}\ Z \end{array} \tag{2.26}$$

Given $\mathbb{K} : X \rightarrowtriangle Y$ and $\mathbb{L} : Y \times X \rightarrowtriangle Z$, the semidirect product is graphically represented by connecting $\mathbb{K}$ and $\mathbb{L}$ and keeping an extra copy

$$\mathbb{K} \odot \mathbb{L} := \text{copy}_X(\mathbb{K} \otimes \text{id}_X)(\text{copy}_Y \otimes \text{id}_X)(\text{id}_Y \otimes \mathbb{L}) \tag{2.27}$$

$=$ 

(2.28)

A space X is identified with the identity kernel $\text{id}^X : X \to \Delta(\mathcal{X})$. A bare wire represents the identity kernel:

$$\text{Id}^X := X \ \text{———}\ X \tag{2.29}$$

Product spaces $X \times Y$ are identified with tensor product of identity kernels $\text{id}^X \otimes \text{id}^Y$. These can be represented either by two parallel wires or by a single wire representing the identity on the product space $X \times Y$:

$$X \times Y \cong \text{Id}^X \otimes \text{Id}^Y := \begin{array}{l} X \text{ — } X \\ Y \text{ — } Y \end{array} \tag{2.30}$$

$$= X \times Y \ \text{———}\ X \times Y \tag{2.31}$$

A kernel $\mathbb{L} : X \to \Delta(\mathcal{Y} \otimes \mathcal{Z})$ can be written using either two parallel output wires or a single output wire, appropriately labeled:

$$X \text{ —}\boxed{\mathbb{L}}\!\begin{array}{l} \text{— } Y \\ \text{— } Z \end{array} \tag{2.32}$$

$$\equiv \tag{2.33}$$

$$X \text{ —}\boxed{\mathbb{L}}\text{— } Y \times Z \tag{2.34}$$

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.

2.3.2 Special maps

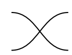
Definition 2.3.1 (Identity map)**.** The identity map $\mathrm{id}_X : X \rightarrowtriangle X$ defined by $(\mathrm{id}_X)(A|x) = \delta_x(A)$ for all $x \in X$, $A \in \mathcal{X}$, is represented by a bare line.

$$\mathrm{id}_X := \; X \text{-} X \tag{2.35}$$

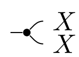
Definition 2.3.2 (Erase map)**.** Given some 1-element set $\{*\}$, the erase map $\mathrm{del}_X : X \rightarrowtriangle \{*\}$ is defined by $(\mathrm{del}_X)(*|x) = 1$ for all $x \in X$. It "discards the input". It looks like a lit fuse:

$$\mathrm{del}_X := \quad -\!\!\ast\; X \tag{2.36}$$

Definition 2.3.3 (Swap map)**.** The swap map $\mathrm{swap}_{X,Y} : X \times Y \rightarrowtriangle Y \times X$ is defined by $(\mathrm{swap}_{X,Y})(A \times B|x,y) = \delta_x(B)\delta_y(A)$ for $(x,y) \in X \times Y$, $A \in \mathcal{X}$ and $B \in \mathcal{Y}$. It swaps two inputs and is represented by crossing wires:

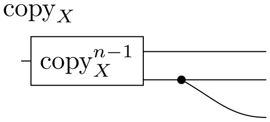
$\mathrm{swap}_{X,Y} :=$  (2.37)

Definition 2.3.4 (Copy map)**.** The copy map $\mathrm{copy}_X : X \rightarrowtriangle X \times X$ is defined by $(\mathrm{copy}_X)(A \times B|x) = \delta_x(A)\delta_x(B)$ for all $x \in X$, $A, B \in \mathcal{X}$. It makes two identical copies of the input, and is drawn as a fork:

$\mathrm{copy}_X := X$  (2.38)

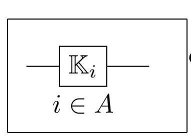
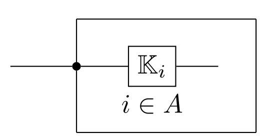
Definition 2.3.5 (n-fold copy map)**.** The n-fold copy map $\mathrm{copy}^n_X : X \rightarrowtriangle X^n$ is given by the recursive definition

$$\mathrm{copy}^1_X = \mathrm{copy}_X \tag{2.39}$$

$\mathrm{copy}^n_X =$  $n > 1$ (2.40)

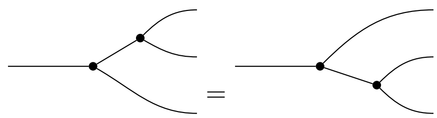
Plates In a string diagram, a plate that is annotated $i \in A$ means the tensor product of the $|A|$ elements that appear inside the plate. A wire crossing from outside a plate boundary to the inside of a plate indicates an $|A|$-fold copy map, which we indicate by placing a dot on the plate boundary. For our purposes, 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 $\mathbb{K}_i : X \rightarrowtriangle Y$ for $i \in A$,

$\bigotimes_{i\in A} \mathbb{K}_i :=$  $\mathrm{copy}^{|A|}_X(\bigotimes_{i\in A} \mathbb{K}_i) \quad :=$  (2.41)

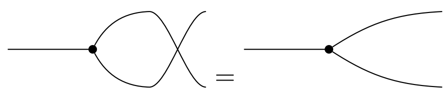
2.3.3 Commutative comonoid axioms

Diagrams in Markov categories satisfy the commutative comonoid axioms.



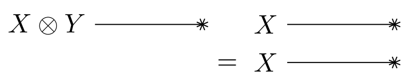
(2.42)

(2.43)

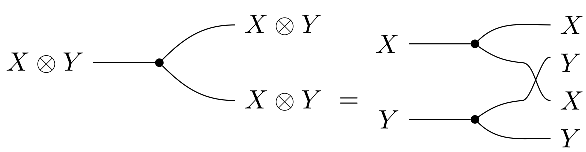


(2.44)

as well as compatibility with the monoidal structure

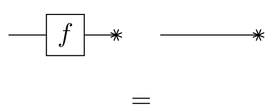


(2.45)



(2.46)

and the naturality of *del*, which means that

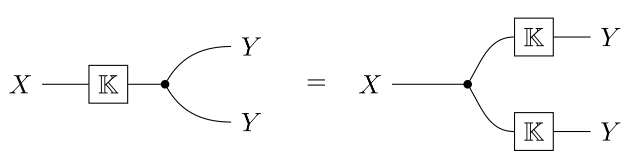


(2.47)

2.3.4 Manipulating String Diagrams

Planar deformations along with the applications of Equations 2.42 through to Equation 2.47 are almost the only rules we have for transforming one string diagram into an equivalent one. One further rule is given by Theorem 2.3.6.

Theorem 2.3.6 (Copy map commutes for deterministic kernels (Fong, 2013))**.** *For* $\mathbb{K} : X \rightarrow\!\!\!\triangleright\, Y$



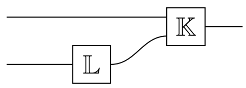
(2.48)

holds iff $\mathbb{K}$ *is deterministic.*

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.

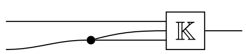
For arbitrary $\mathbb{K} : X \times Y \rightarrowtriangle Z$, $\mathbb{L} : W \rightarrowtriangle Y$



$$
\begin{aligned}
&= (\mathrm{id}_X \otimes \mathbb{L})\mathbb{K} && (2.49)\\
[(\mathrm{id}_X \otimes \mathbb{L})\mathbb{K}](A|x,w) &= \int_Y \int_X \mathbb{K}(A|x',y')\mathbb{L}(\mathrm{d}y'|w)\delta_x(\mathrm{d}x') && (2.50)\\
&= \int_Y \mathbb{K}(A|x,y')\mathbb{L}(dy'|w) && (2.51)
\end{aligned}
$$

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

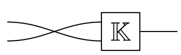
For arbitrary $\mathbb{K} : X \times Y \times Y \rightarrowtriangle Z$:



$$
\begin{aligned}
&= (\mathrm{id}_X \otimes \mathrm{copy}_Y)\mathbb{K} && (2.52)\\
[(\mathrm{id}_X \otimes \mathrm{copy}_Y)\mathbb{K}](A|x,y) &= \int_Y \int_Y \mathbb{K}(A|x,y',y'')\delta_y(\mathrm{d}y')\delta_y(\mathrm{d}y'') && (2.53)\\
&= \mathbb{K}(A|x,y,y) && (2.54)
\end{aligned}
$$

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

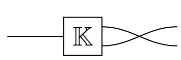
For arbitrary $\mathbb{K} : X \times Y \rightarrowtriangle Z$



$$
\begin{aligned}
&= \mathrm{swap}_{YX}\mathbb{K} && (2.55)\\
(\mathrm{swap}_{YX}\mathbb{K})(A|y,x) &= \int_{X\times Y} \mathbb{K}(A|x',y')\delta_y(\mathrm{d}y')\delta_x(\mathrm{d}x') && (2.56)\\
&= \mathbb{K}(A|x,y) && (2.57)
\end{aligned}
$$

The swap map before a kernel switches the input arguments.

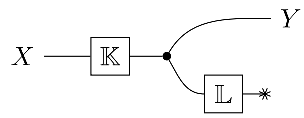
For arbitrary $\mathbb{K} : X \rightarrowtriangle Y \times Z$



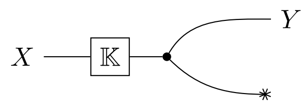
$$
\begin{aligned}
&= \mathbb{K}\mathrm{swap}_{YZ} && (2.58)\\
(\mathbb{K}\mathrm{swap}_{YZ})(A\times B|x) &= \int_{Y\times Z} \delta_y(B)\delta_z(A)\mathbb{K}(\mathrm{d}y \times \mathrm{d}z|x) && (2.59)\\
&= \int_{B\times A} \mathbb{K}(\mathrm{d}y \times \mathrm{d}z|x) && (2.60)\\
&= \mathbb{K}(B \times A|x) && (2.61)
\end{aligned}
$$

Given $\mathbb{K} : X \rightarrowtriangle Y$ and $\mathbb{L} : Y \rightarrowtriangle Z$:

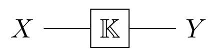
$$(\mathbb{K} \odot \mathbb{L})(\mathrm{id}_Y \otimes \mathrm{del}_Z) = \tag{2.62}$$



$$= \qquad \text{by Eq. 2.47} \tag{2.63}$$



$$= \qquad \text{by Eq. 2.43} \tag{2.64}$$



Thus the action of the del map is to marginalise over the deleted wire. With integrals, we can write

$$(\mathbb{K} \odot \mathbb{L})(\mathrm{id}_Y \otimes \mathrm{del}_Z)(A \times \{*\}|x) = \int_Y \int_{\{*\}} \delta_y(A)\delta_*(\{*\})\mathbb{L}(\mathrm{d}z|y)\mathbb{K}(\mathrm{d}y|x) \tag{2.65}$$

$$= \int_A \mathbb{K}(\mathrm{d}y|x) \tag{2.66}$$

$$= \mathbb{K}(A|x) \tag{2.67}$$

2.4 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 2.4.1 (Probability set)**.** A probability set $\mathbb{P}_C$ on $(\Omega, \mathcal{F})$ is a collection of probability measures on $(\Omega, \mathcal{F})$. In other words it is a subset of $\mathcal{P}(\Delta(\Omega))$, where $\mathcal{P}$ indicates the power set.

Given a probability set $\mathbb{P}_C$, we define marginal and conditional probabilities as probability measures and Markov kernels that satisfy Definitions 2.2.15 and 2.2.16 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 2.4.7) with respect to probability sets.

Definition 2.4.2 (Marginal probability with respect to a probability set)**.** Given a sample space $(\Omega, \mathcal{F})$, a variable $\mathsf{X} : \Omega \to X$ and a probability set $\mathbb{P}_C$, the marginal distribution $\mathbb{P}_C^{\mathsf{X}} = \mathbb{P}_\alpha^{\mathsf{X}}$ for any $\mathbb{P}_\alpha \in \mathbb{P}_C$ if a distribution satisfying this condition exists. Otherwise, it is undefined.

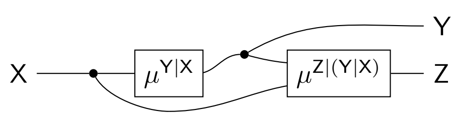
Definition 2.4.3 (Uniform conditional distribution)**.** Given a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$ and a probability set $\mathbb{P}_C$, a uniform conditional distribution $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ is any Markov kernel $X \rightarrowtriangle Y$ such that $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ is an $\mathsf{Y}|\mathsf{X}$ conditional probability of $\mathbb{P}_\alpha$ for all $\mathbb{P}_\alpha \in \mathbb{P}_C$. If no such Markov kernel exists, $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ is undefined.

Given a conditional distribution $\mu^{\mathsf{ZY}|\mathsf{X}}$ we can define a higher order conditional $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$, which is a version of $\mu^{\mathsf{Z}|\mathsf{XY}}$. 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 $\mathbb{P}_C^{\mathsf{ZY}|\mathsf{X}}$ then it also has a uniform conditional $\mathbb{P}_C^{\mathsf{Z}|\mathsf{XY}}$ (Theorems 2.4.31 and 2.4.33). Given $\mu^{\mathsf{XY}|\mathsf{Z}}$ and $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ standard measurable, it has recently been proven that a higher order conditional $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ exists Bogachev and Malofeev (2020), Theorem 3.5.

Definition 2.4.4 (Higher order conditionals)**.** Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and $\mathsf{Z} : \Omega \to Z$, a higher order conditional $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})} : X \times Y \to Z$ is any Markov kernel such that, for some $\mu^{\mathsf{Y}|\mathsf{X}}$,

$$\mu^{\mathsf{ZY}|\mathsf{X}}(B \times C|x) = \int_B \mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}(C|x, y)\mu^{\mathsf{Y}|\mathsf{X}}(dy|x) \tag{2.68}$$

$$\iff \tag{2.69}$$



$$\mu^{\mathsf{ZY}|\mathsf{X}} = \tag{2.70}$$

Definition 2.4.5 (Uniform higher order conditional)**.** Given a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_C$, if $\mathbb{P}_C^{\mathsf{ZY}|\mathsf{X}}$ exists then a uniform higher order conditional $\mathbb{P}_C^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ is any Markov kernel $X \times Y \rightarrowtriangle Z$ that is a higher order conditional of some version of $\mathbb{P}_C^{\mathsf{ZY}|\mathsf{X}}$. If no $\mathbb{P}_C^{\mathsf{ZY}|\mathsf{X}}$ exists, $\mathbb{P}_C^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ is undefined.

Definition 2.4.6 (Almost sure equality)**.** Two Markov kernels $\mathbb{K} : X \rightarrowtriangle Y$ and $\mathbb{L} : X \rightarrowtriangle Y$ are $\mathbb{P}_C, \mathsf{X}, \mathsf{Y}$-almost surely equal if for all $A \in \mathcal{X}$, $B \in \mathcal{Y}$, $\alpha \in C$

$$\int_A \mathbb{K}(B|x)\mathbb{P}_\alpha^{\mathsf{X}}(\mathrm{d}x) = \int_A \mathbb{L}(B|x)\mathbb{P}_\alpha^{\mathsf{X}}(\mathrm{d}x) \tag{2.71}$$

we write this as $\mathbb{K} \overset{\mathbb{P}_C}{\cong} \mathbb{L}$, 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 $\mathbb{P}_\alpha^{\mathsf{X}}$ for all $\alpha \in C$.

2.4.1 Almost sure equality

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

Definition 2.4.7 (Almost sure equality)**.** Two Markov kernels $\mathbb{K} : X \rightarrowtriangle Y$ and $\mathbb{L} : X \rightarrowtriangle Y$ are almost surely equal $\overset{\mathbb{P}_C}{\cong}$ with respect to a probability set $\mathbb{P}_C$ and variable $\mathsf{X} : \Omega \to X$ if for all $\mathbb{P}_\alpha \in \mathbb{P}_C$,

$$\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{K} = \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{L} \tag{2.72}$$

Lemma 2.4.8 (Uniform conditional distributions are almost surely equal)**.** *If $\mathbb{K} : X \rightarrowtriangle Y$ and $\mathbb{L} : X \rightarrowtriangle Y$ are both versions of $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ then $\mathbb{K} \overset{\mathbb{P}_C}{\cong} \mathbb{L}$*

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

$$\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{K} = \mathbb{P}_\alpha^{\mathsf{XY}} \tag{2.73}$$

$$= \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{L} \tag{2.74}$$

□

Lemma 2.4.9 (Substitution of almost surely equal Markov kernels)**.** *Given $\mathbb{P}_C$, if $\mathbb{K} : X \times Y \rightarrowtriangle Z$ and $\mathbb{L} : X \times Y \rightarrowtriangle Z$ are almost surely equal $\mathbb{K} \overset{\mathbb{P}_C}{\cong} \mathbb{L}$, then for any $\mathbb{P}_\alpha \in \mathbb{P}_C$*

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{K} \overset{\mathbb{P}_C}{\cong} \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{L} \tag{2.75}$$

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

$$\mathbb{P}_\alpha^{\mathsf{XY}} \odot \mathbb{K} \overset{\mathbb{P}_C}{\cong} (\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}) \odot \mathbb{K} \tag{2.76}$$

$$\overset{\mathbb{P}_C}{\cong} \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{K}) \tag{2.77}$$

$$\overset{\mathbb{P}_C}{\cong} \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{L}) \tag{2.78}$$

□

Theorem 2.4.10 (Semidirect product of uniform conditional distributions is a joint uniform conditional distribution)**.** *Given a probability set $\mathbb{P}_C$ on $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and uniform conditional distributions $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ and $\mathbb{P}_C^{\mathsf{Z}|\mathsf{XY}}$, then $\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}$ exists and is equal to*

$$\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}} \overset{\mathbb{P}_C}{\cong} \mathbb{P}_C^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_C^{\mathsf{Z}|\mathsf{XY}} \tag{2.79}$$

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

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{XYZ}} &= \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_\alpha^{\mathsf{YZ}|\mathsf{X}} && (2.80)\\
&= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_\alpha^{\mathsf{Z}|\mathsf{YX}}) && (2.81)\\
&= \mathbb{P}_\alpha^{\mathsf{X}} \odot (\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{P}_C^{\mathsf{Z}|\mathsf{YX}}) && (2.82)
\end{aligned}$$

□

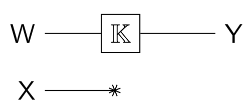
2.4.2 Extended conditional independence

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 2.4.12).

Definition 2.4.11 (Conditional independence)**.** For a *probability model* $\mathbb{P}_\alpha$ and variables $\mathsf{A}, \mathsf{B}, \mathsf{Z}$, we say B is conditionally independent of A given C, written $\mathsf{B} \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{A}|\mathsf{C}$, if

$$\mathbb{P}^{\mathsf{Y}|\mathsf{WX}} \overset{\mathbb{P}}{\cong} \quad (2.83)$$

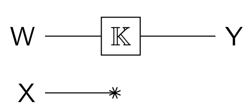


$$\iff \mathbb{P}^{\mathsf{Y}|\mathsf{WX}}(A|w,x) \overset{\mathbb{P}}{\cong} \mathbb{K}(A|w) \qquad \forall A \in \mathcal{Y} \qquad (2.84)$$

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

Theorem 2.4.12. *Given standard measurable* $(\Omega, \mathcal{F})$*, a probability model* $\mathbb{P}$ *and variables* $\mathsf{W} : \Omega \to W$*,* $\mathsf{X} : \Omega \to X$*,* $\mathsf{Y} : \Omega \to Y$*,* $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}} \mathsf{X}|\mathsf{W}$ *if and only if there exists some version of* $\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}$ *and* $\mathbb{K} : W \rightarrowtriangle Y$ *such that*

$$\mathbb{P}^{\mathsf{Y}|\mathsf{WX}} \overset{\mathbb{P}}{\cong} \quad (2.85)$$



$$\iff \mathbb{P}^{\mathsf{Y}|\mathsf{WX}}(A|w,x) \overset{\mathbb{P}}{\cong} \mathbb{K}(A|w) \qquad \forall A \in \mathcal{Y} \qquad (2.86)$$

Proof. See Cho and Jacobs (2019). □

Extended conditional independence as introduced by Constantinou and Dawid (2017) is defined in terms of "nonstochastic variables" on the choice set C. A nonstochastic variable is essentially a variable defined on C rather than on the sample space Ω

Definition 2.4.13 (Nonstochastic variable)**.** Given a sample space $(\Omega, \mathcal{F})$, a choice set $(C, \mathcal{C})$, a codomain $(X, \mathcal{X})$ and a probability set $\mathbb{P}_C$, a nonstochastic variable is a measurable function $\phi : C \to X$.

In particular, we want to consider *complementary* nonstochastic variable - that is, pairs of nonstochastic variables ϕ and ξ such that the sequence (ϕ, ξ) is invertible. For example, if $\phi := \mathrm{idf}_C$, then

Definition 2.4.14 (Complementary nonstochastic variables)**.** A pair of nonstochasti variables ϕ and ξ are complementary if (ϕ, ξ) is invertible.

Nxample 2.4.15. The letters ϕ and ξ are used to represent complementary nonstochastic variables.

Unlike Constantinou and Dawid (2017), we limit ourselves to a definition of extended conditional independence where regular uniform conditional probabilities exist. Our definition is otherwise identical.

Definition 2.4.16 (Extended conditional independence)**.** Given a probability set $\mathbb{P}_C$, variables X, Y and Z and complementary nonstochastic variables ϕ and ξ, the extended conditional independence $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{X}\phi|\mathsf{Z}\xi$ holds if for each $a \in \xi(C)$, $\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}_{\xi^{-1}(a)}$ and $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}_{\xi^{-1}(a)}$ exist and

$$\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}_{\xi^{-1}(a)} \overset{\mathbb{P}_{\xi^{-1}(a)}}{\cong} \begin{array}{l} \mathsf{Z} \longrightarrow \boxed{\mathbb{P}_C^{\mathsf{Y}|\mathsf{Z}}} \longrightarrow \mathsf{Y} \\ \mathsf{X} \longrightarrow * \end{array} \tag{2.87}$$

$$\iff \tag{2.88}$$

$$\mathbb{P}^{\mathsf{Y}|\mathsf{XZ}}_{\xi^{-1}(a)}(A|x,z) \overset{\mathbb{P}_{\xi^{-1}(a)}}{\cong} \mathbb{P}^{\mathsf{Y}|\mathsf{Z}}_{\xi^{-1}(a)}(A|z) \qquad \forall A \in \mathcal{Y}, (x,z) \in X \times Z \tag{2.89}$$

Very often, we consider a particular kind of extended conditional independence that does not explicitly make use of nonstochastic variables. We call this *uniform conditional independence.*

Definition 2.4.17 (Uniform conditional independence)**.** Given a probability set $\mathbb{P}_C$ and variables X, Y and Z, the uniform conditional independence $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{X}C|\mathsf{Z}$ holds if $\mathbb{P}_C^{\mathsf{Y}|\mathsf{XZ}}$ and $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ exist and

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{XZ}} \overset{\mathbb{P}_C}{\cong} \begin{array}{l} \mathsf{Z} \longrightarrow \boxed{\mathbb{P}_C^{\mathsf{Y}|\mathsf{Z}}} \longrightarrow \mathsf{Y} \\ \mathsf{X} \longrightarrow * \end{array} \tag{2.90}$$

$$\iff \tag{2.91}$$

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{XZ}}(A|x,z) \overset{\mathbb{P}_C}{\cong} \mathbb{P}_C^{\mathsf{Y}|\mathsf{Z}}(A|z) \qquad \forall A \in \mathcal{Y}, (x,z) \in X \times Z \tag{2.92}$$

For countable sets C (which, recall, is an assumption we generally accept), as shown by Constantinou and Dawid (2017) we can reason with collections of extended conditional independence statements as if they were regular conditional independence statements, with the provision that a complementary pair of nonstochastic variables must appear either side of the "|" symbol.

1. Symmetry: $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}\phi|\mathsf{Z}\xi$ iff $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{X}\phi|\mathsf{Z}\xi$
2. $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}C|\mathsf{Y}C$
3. Decomposition: $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}\phi|\mathsf{W}\xi$ and $\mathsf{Z} \preccurlyeq \mathsf{Y}$ implies $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Z}\phi|\mathsf{W}\xi$
4. Weak union:
   (a) $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}\phi|\mathsf{W}\xi$ and $\mathsf{Z} \preccurlyeq \mathsf{Y}$ implies $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}\phi|(\mathsf{Z},\mathsf{W})\xi$
   (b) $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}\phi|\mathsf{W}\xi$ and $\lambda \preccurlyeq \phi$ implies $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}\phi|(\mathsf{Z},\mathsf{W})(\xi,\lambda)$
5. Contraction: $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Z}\phi|\mathsf{W}\xi$ and $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}\phi|(\mathsf{Z},\mathsf{W})\xi$ implies $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y},\mathsf{Z})\phi|\mathsf{W}\xi$

The following forms of these properties are often used here:

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

2.4.3 Examples

Example 2.4.18 (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 $\mathcal{S}_\alpha$ is modeled by a probability distribution $\mathbb{P}_\alpha$ on a shared sample space $(\Omega, \mathcal{F})$ such that we have an observable "choice" variable $(\mathsf{D}, \mathsf{D}\circ\mathcal{S}_\alpha)$ where $\mathsf{D}\circ\mathcal{S}_\alpha$ always yields α.

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

$$\mathbb{P}_\alpha^{\mathsf{YC}}(A \times B) = \mathbb{P}_\alpha(A)\delta_\alpha(B) \tag{2.93}$$

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

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}} = \mathbb{P}_\alpha^{\mathsf{Y}} \tag{2.94}$$

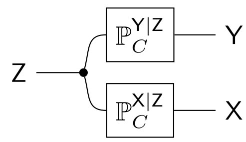
Thus $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ exists and

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}(A|\alpha) = \mathbb{P}_\alpha^{\mathsf{Y}}(A) \qquad\qquad \forall A \in \mathcal{Y}, \alpha \in C \tag{2.95}$$

Because only deterministic marginals $\mathbb{P}_\alpha^{\mathsf{D}}$ are available, for every $\alpha \in C$ we have $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{D}$. 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 2.4.19 (Uniform conditional independence representation)**.** *Given a probability set* $\mathbb{P}_C$ *with a uniform conditional probability* $\mathbb{P}_C^{\mathsf{XY|Z}}$*,*

$$\mathbb{P}_C^{\mathsf{XY|Z}} \overset{\mathbb{P}_C}{\cong} \tag{2.96}$$



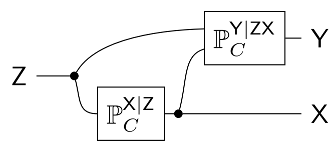
$$\iff \tag{2.97}$$

$$\mathbb{P}_C^{\mathsf{XY|Z}}(A \times B|z) \overset{\mathbb{P}_C}{\cong} \mathbb{P}_C^{\mathsf{X|Z}}(A|z)\mathbb{P}_C^{\mathsf{Y|Z}}(B|z) \qquad \forall A \in \mathcal{X}, B \in \mathcal{Y}, z \in Z \tag{2.98}$$

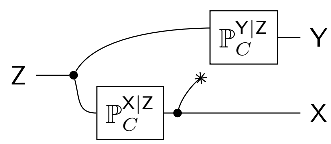
if and only if $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{X}C|\mathsf{Z}$

Proof. If: By Theorem 2.4.33

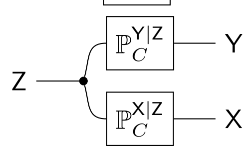
$$\mathbb{P}_C^{\mathsf{XY|Z}} = \tag{2.99}$$



$$\overset{\mathbb{P}_C}{\cong} \tag{2.100}$$

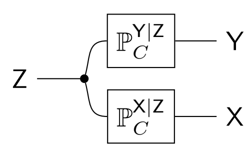


$$= \tag{2.101}$$



Only if: Suppose

$$\mathbb{P}_C^{\mathsf{XY|Z}} \overset{\mathbb{P}_C}{\cong} \tag{2.102}$$



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

$$\mathbb{P}_C^{\mathsf{Y|XZ}}(B|x,z) > \mathbb{P}_C^{\mathsf{Y|Z}}(B|z) \qquad \forall (x,z) \in A \times C \tag{2.103}$$

then

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{XYZZ}}(A \times B \times C) &= \int_{A\times C} \mathbb{P}_C^{\mathsf{Y|XZ}}(B|x,z)\mathbb{P}_C^{\mathsf{X|Z}}(\mathrm{dx}|z)\mathbb{P}_\alpha^{\mathsf{Z}}(\mathrm{dz}) && (2.104)\\
&> \int_{A\times C} \mathbb{P}_C^{\mathsf{Y|X}}(B|z)\mathbb{P}_C^{\mathsf{X|Z}}(\mathrm{dx}|z)\mathbb{P}_\alpha^{\mathsf{Z}}(\mathrm{dz}) && (2.105)\\
&= \int_C \mathbb{P}_C^{\mathsf{XY|X}}(A \times B|z)\mathbb{P}_\alpha^{\mathsf{Z}}(\mathrm{dz}) && (2.106)\\
&= \mathbb{P}_\alpha^{\mathsf{XYZZ}}(A \times B \times C) && (2.107)
\end{aligned}$$

a contradiction. An analogous argument follows if we replace "$>$" with "$<$" in Eq. 2.103. $\square$

2.4.4 Maximal probability sets and valid conditionals

So far, we have been implicitly supposing that we first set up a probability set and from that set we may sometimes derive uniform conditional probabilities, extended conditional independences and so forth. However, sometimes we want to work backwards: start with a collection of uniform conditional probabilities, and work with the probability set implicitly defined by this collection. For example, when we have a Causal Bayesian Network, the collection of operations of the form "do($\mathsf{X} = x$)" specify a probability set by a collection of uniform conditional probabilities on variables other than X, along with marginal probabilities of X. Specifically:

$$\mathbb{P}_{\mathsf{X}=x}^{\mathsf{Y}|\mathrm{Pa}(\mathsf{Y})} = \begin{cases} \mathbb{P}_{\mathrm{obs}}^{\mathsf{Y}|\mathrm{Pa}(\mathsf{Y})} & \mathsf{Y} \text{ is a causal variable and not equal to } \mathsf{X} \\ \delta_x & \mathsf{Y} = \mathsf{X} \end{cases} \tag{2.108}$$

The qualification "Y is a causal variable" is usually not an explicit condition for causal Bayesian networks, but it is an important one. For example, $2\mathsf{X}$ is not equal to X, but we cannot define a causal Bayesian network where both X and $2\mathsf{X}$ are causal variables, see Example 2.4.27.

When working backwards like this, we can run into a couple of problems: we may end up with a probability set where some probabilities are non-unique, or we might inadvertently define an empty probability set. *Validity* is a condition that can ensure that we at least avoid the second problem.

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 $\mathsf{X} \longrightarrow \mathsf{Y}$ and some observational probability distribution $\mathbb{P}^{\mathsf{XY}} \in \Delta(X \times Y)$. Using the standard notion of "hard interventions on X", this model induces a probability set which we could informally describe as the set $\mathbb{P}_\square := \{\mathbb{P}_a^{\mathsf{XY}} | a \in X \cup \{*\}\}$ where $*$ is a special element corresponding to the observational setting. The graph $\mathsf{X} \longrightarrow \mathsf{Y}$ implies the existence of the uniform conditional probability $\mathbb{P}_\square^{\mathsf{Y}|\mathsf{X}}$ under the nominated set of interventions, while the usual rules of hard interventions imply that $\mathbb{P}_a^{\mathsf{X}} = \delta_a$ 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 $\mathsf{X} = (\mathsf{Z}, \mathsf{Z})$ for $\mathsf{Z} := \mathrm{id}_\Omega$ and any measure $\kappa \in \Delta(\{0, 1\}^2)$ such that $\kappa(\{1\} \times \{0\}) > 0$. Note that $\mathsf{X}^{-1}(\{1\} \times \{0\}) =$

$\mathsf{Z}^{-1}(\{1\}) \cap \mathsf{Z}^{-1}(\{0\}) = \emptyset$. Thus for any probability measure $\mu \in \Delta(\{0,1\})$, $\mu^{\mathsf{X}}(\{1\} \times \{0\}) = \mu(\emptyset) = 0$ and so κ cannot be the marginal distribution of X for any base measure at all.

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

Definition 2.4.21 (Valid conditional)**.** Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ a $\mathsf{Y}|\mathsf{X}$-*valid conditional probability* is a Markov kernel $\mathbb{L} : X \rightarrow Y$ that assigns probability 0 to impossible events, unless the argument itself corresponds to an impossible event:

$$\forall B \in \mathcal{Y}, x \in X : (\mathsf{X}, \mathsf{Y}) \bowtie \{x\} \times B = \emptyset \implies (\mathbb{L}(B|x) = 0) \vee (\mathsf{X} \bowtie \{x\} = \emptyset) \tag{2.109}$$

When a probability distribution is interpreted as a Markov kernel, both of these definitions agree.

Theorem 2.4.22 (Equivalence of validity definitions)**.** *Given $\mathsf{X} : \Omega \to X$, with Ω and X standard measurable, a probability measure $\mathbb{P}^{\mathsf{X}} \in \Delta(X)$ is valid if and only if the conditional $\mathbb{P}^{\mathsf{X}|*} := * \mapsto \mathbb{P}^{\mathsf{X}}$ is valid.*

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

$$\forall A \in \mathcal{X} : \mathsf{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathsf{X}|*}(A|*) = 0 \tag{2.110}$$

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

$$\forall A \in \mathcal{X} : \mathsf{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathsf{X}}(A) = 0 \tag{2.111}$$

□

Conditionals can be used to define *maximal probability sets*, which is the set of all probability distributions with those conditionals.

Definition 2.4.23 (Maximal probability set)**.** Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and a $\mathsf{Y}|\mathsf{X}$-valid conditional probability $\mathbb{L} : X \rightarrowtriangle Y$ the maximal probability set $\mathbb{P}_C$ 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 $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}$.

Theorem 2.4.24 shows that the semidirect product of any pair of valid conditional probabilities is itself a valid conditional. Suppose we have some collection of $\mathsf{X}_i|\mathsf{X}_{[i-1]}$-valid conditionals $\{\mathbb{P}_i^{\mathsf{X}_i|\mathsf{X}_{[i-1]}}|i \in [n]\}$; then recursively taking the semidirect product $\mathbb{M} := \mathbb{P}_1^{\mathsf{X}_1} \odot (\mathbb{P}_2^{\mathsf{X}_2|\mathsf{X}_1} \odot ...)$ yields a $\mathsf{X}_{[n]}$ valid distribution. Furthermore, the maximal probability set associated with $\mathbb{M}$ is nonempty.

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

Note that validity is not a necessary condition for a conditional to define a non-empty probability set. Given some $\mathbb{K} : X \rightarrowtriangle Y$, $\mathbb{K}$ might be an invalid conditional on if every value of X is considered, but it might be valid on some subset of X. A marginal of X that assigns measure 0 to the subset of X where $\mathbb{K}$ is invalid can still define a valid distribution when combined with $\mathbb{K}$. On the other hand, if $\mathbb{K}$ is required to combine with arbitrary valid marginals of X, then the validity of $\mathbb{K}$ is necessary (Theorem 2.4.26).

Theorem 2.4.24 (Semidirect product of valid conditional distributions is valid)**.** *Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ (all spaces standard measurable) and any valid candidate conditional $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ and $\mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}$, $\mathbb{P}^{\mathsf{Y}|\mathsf{X}} \odot \mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}$ is also a valid candidate conditional.*

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

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

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

$$\mathbb{R}^{\mathsf{YZ}|\mathsf{X}}(B \times C|x) = \int_B \mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}(C|y,x)\mathbb{P}^{\mathsf{Y}|\mathsf{X}}(dy|x) \tag{2.112}$$
$$\leq \mathbb{P}^{\mathsf{Y}|\mathsf{X}}(B|x) \tag{2.113}$$
$$= 0 \tag{2.114}$$

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

$$\mathbb{R}^{\mathsf{YZ}|\mathsf{X}}(B \times C|x) = \int_B \mathbb{Q}^{\mathsf{Z}|\mathsf{YX}}(C|y,x)\mathbb{P}^{\mathsf{Y}|\mathsf{X}}(dy|x) \tag{2.115}$$
$$= 0 \tag{2.116}$$

□

Corollary 2.4.25 (Valid conditionals are validly extendable to valid distributions)**.** *Given Ω, $\mathsf{U} : \Omega \to U$, $\mathsf{W} : \Omega \to W$ and a valid conditional $\mathbb{T}^{\mathsf{W}|\mathsf{U}}$, then for any valid conditional $\mathbb{V}^{\mathsf{U}}$, $\mathbb{V}^{\mathsf{U}} \odot \mathbb{T}^{\mathsf{W}|\mathsf{U}}$ is a valid probability.*

Proof. Applying Lemma 2.4.24 choosing $\mathsf{X} = *$, $\mathsf{Y} = \mathsf{U}$, $\mathsf{Z} = \mathsf{W}$ and $\mathbb{P}^{\mathsf{Y}|\mathsf{X}} = \mathbb{V}^{\mathsf{U}|*}$ and $\mathbb{Q}^{\mathsf{Z}|\mathsf{YX}} = \mathbb{T}^{\mathsf{W}|\mathsf{U}*}$ we have $\mathbb{R}^{WU|*} := \mathbb{V}^{\mathsf{U}|*} \odot \mathbb{T}^{\mathsf{W}|\mathsf{U}*}$ is a valid conditional probability. Then $\mathbb{R}^{\mathsf{WU}} \cong \mathbb{R}^{\mathsf{WU}|*}$ is valid by Theorem 2.4.22. □

Theorem 2.4.26 (Validity of conditional probabilities)**.** *Suppose we have Ω, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, with Ω, X, Y discrete. A conditional $\mathbb{T}^{\mathsf{Y}|\mathsf{X}}$ is valid if and only if for all valid distributions $\mathbb{V}^{\mathsf{X}}$, $\mathbb{V}^{\mathsf{X}} \odot \mathbb{T}^{\mathsf{Y}|\mathsf{X}}$ is also a valid distribution.*

Proof. If: this follows directly from Corollary 2.4.25.

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

$$(\mathbb{V}^{\mathsf{X}} \odot \mathbb{T}^{\mathsf{Y}|\mathsf{X}})(x, y) = \mathbb{T}^{\mathsf{Y}|\mathsf{X}}(y|x)\mathbb{V}^{\mathsf{X}}(x) \tag{2.117}$$

$$= \mathbb{T}^{\mathsf{Y}|\mathsf{X}}(y|x) \tag{2.118}$$

$$> 0 \tag{2.119}$$

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

Example 2.4.27. Body mass index is defined as a person's weight divided by the square of their height. Suppose we have a measurement process $\mathcal{S} = (\mathcal{W}, \mathcal{H})$ and $\mathcal{B} = \frac{\mathcal{W}}{\mathcal{H}^2}$ - i.e. we figure out someone's body mass index first by measuring both their height and weight, and then passing the result through a function that divides the second by the square of the first. Thus, given the random variables W, H modelling $\mathcal{W}, \mathcal{H}$, $\mathcal{B}$ is the function given by $\mathsf{B} = \frac{\mathsf{W}}{\mathsf{H}^2}$.

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

$$\mathbb{P}_\square^{\mathsf{B}|\mathsf{WHC}} = \begin{array}{l} \mathsf{H} \longrightarrow * \\ \mathsf{C} \longrightarrow \mathsf{B} \\ \mathsf{W} \longrightarrow * \end{array} \tag{2.120}$$

Then pick some $w, h, x \in \mathbb{R}$ such that $\frac{w}{h^2} \neq x$ and $(\mathsf{W}, \mathsf{H}) \bowtie (w, h) \neq \emptyset$ (which is to say, our measurement procedure could potentially yield (w, h) for a person's height and weight). We have $\mathbb{P}_\square^{\mathsf{B}|\mathsf{WHC}}(\{x\}|w, h, x) = 1$, but

$$(\mathsf{B}, \mathsf{W}, \mathsf{H}) \bowtie \{(x, w, h)\} = \{\omega | (\mathsf{W}, \mathsf{H})(\omega) = (w, h), \mathsf{B}(\omega) = \frac{w}{h^2}\} \tag{2.121}$$

$$= \emptyset \tag{2.122}$$

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

Validity rules out conditional probabilities like 2.120. 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 2.120 would be used to evaluate the causal effect of body mass index in the causal diagram found in Shahar (2009).

2.4.5 Existence of conditional probabilities

Lemma 2.4.28 (Conditional pushforward)**.** *Suppose we have a sample space* $(\Omega, \mathcal{F})$*, variables* $\mathsf{X} : \Omega \to X$ *and* $\mathsf{Y} : \Omega \to Y$*,* $\mathsf{Z} : \Omega \to Z$ *and a probability set* $\mathbb{P}_C$ *with conditional* $\mathbb{P}_C^{\mathsf{X}|\mathsf{Y}}$ *such that* $\mathsf{Z} = f \circ \mathsf{Y}$ *for some* $f : Y \to Z$*. Then there exists a conditional probability* $\mathbb{P}_C^{\mathsf{Z}|\mathsf{X}} = \mathbb{P}_C^{\mathsf{Y}|\mathsf{X}} \mathbb{F}_f$*.*

Proof. Note that $(\mathsf{X}, \mathsf{Z}) = (\mathrm{id}_X \otimes f) \circ (\mathsf{X}, \mathsf{Y})$. Thus, by Lemma 2.2.21, for any $\mathbb{P}_\alpha \in \mathbb{P}_C$

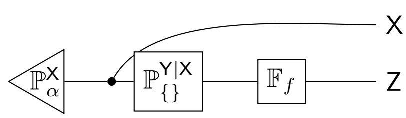
$$\mathbb{P}_\alpha^{\mathsf{XZ}} = \mathbb{P}_\alpha^{\mathsf{XY}} \mathbb{F}_{\mathrm{id}_X \otimes f} \tag{2.123}$$

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

$$\begin{aligned}
\mathbb{F}_{\mathrm{id}_X \otimes f}(A \times B|x, y) &= \delta_x(A)\delta_{f(y)}(B) && (2.124)\\
&= \mathbb{F}_{\mathrm{id}_X}(A|x) \otimes \mathbb{F}_f(B|y) && (2.125)\\
\implies \mathbb{F}_{\mathrm{id}_X \otimes f} &= \mathbb{F}_{\mathrm{id}_X} \otimes \mathbb{F}_f && (2.126)
\end{aligned}$$

Thus

$$\mathbb{P}_\alpha^{\mathsf{XZ}} = (\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_C^{\mathsf{Y}|\mathsf{X}})\mathbb{F}_{\mathrm{id}_X} \otimes \mathbb{F}_f \tag{2.127}$$



$$= \tag{2.128}$$

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

The following theorem is a standard result in many probability texts. In this work, the measurable spaces considered will all be standard measurable and so Theorem 2.4.29 always applies. We will simply assume that conditional probabilities exist, and avoid referencing this theorem every time.

Theorem 2.4.29 (Existence of regular conditionals)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$ with Y standard measurable and a probability model $\mathbb{P}_\alpha$ on $(\Omega, \mathcal{F})$. Then there exists a conditional $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}$.*

Proof. Çinlar (2011, Theorem 2.18) □

The following theorem was proved by Bogachev and Malofeev (2020).

Theorem 2.4.30. *Given a Borel measurable map $m : X \rightarrow Y \times Z$ let $\Pi_Y : Y \times Z \to Y$ be the projection onto Y. Then there exists a Borel measurable map $n : X \times Y \rightarrow Y \times Z$ such that*

$$n(\Pi_y^{-1}(y)|x, y) = 1 \tag{2.129}$$

$$m(\mathsf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x, y) m\mathbb{F}_{\Pi_Y}(dy|x) \quad \forall A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z} \tag{2.130}$$

Proof. Bogachev and Malofeev (2020, Theorem 3.5) □

The following corollary implies that, given a uniform conditional, higher order conditionals can generically be found for probability sets.

Corollary 2.4.31 (Existence of higher order conditionals with respect to probability sets)**.** *Take a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_C$ with uniform conditional distribution $\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}$, and Y and Z standard measurable. Then there exists a higher order uniform conditional $\mathbb{P}_C^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$.*

Proof. Take $\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}$ to be the Borel measurable map m from Theorem 2.4.30, and note that $\Pi_Y \circ (\mathsf{Y}, \mathsf{Z}) = \mathsf{Y}$. Then equation 2.130 implies for all $A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z}$ there is some n such that

$$\begin{aligned}
\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}(\mathsf{Y}^{-1}(A) \cap B|x) &= \int_A n(B|x, y) \mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}} \mathbb{F}_{\Pi_Y}(dy|x) &(2.131)\\
&= \int_A n(B|x, y) \mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}(dy|x) &(2.132)
\end{aligned}$$

where Equation 2.132 follows from Lemma 2.4.28.

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

$$\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}(\mathsf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x, y) \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}(dy|x) \tag{2.133}$$

which implies n is a version of $\mathbb{P}_C^{\mathsf{YZ}|(\mathsf{Y}|\mathsf{X})}$. By Lemma 2.4.28, $n\mathbb{F}_{\Pi_Y}$ is a version of $\mathbb{P}_C^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$. □

We might be motivated to ask whether the higher order conditionals in Theorem 2.4.31 can be chosen to be valid. Despite Lemma 2.4.32 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 2.129 makes n proper with respect to the "wrong" sample space $(Y \times Z, \mathcal{Y} \otimes \mathcal{Z})$ while what we would need is a proper conditional probability with respect to $(\Omega, \mathcal{F})$.

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

Lemma 2.4.32. *Given a probability space $(\mu, \Omega, \mathcal{F})$ and variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, if there is a regular proper conditional probability $\mu^{|\mathsf{X}} : X \rightarrowtriangle \Omega$ then there is a valid conditional distribution $\mu^{\mathsf{Y}|\mathsf{X}}$.*

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

Defining $\mathsf{O} := \mathrm{id}_\Omega$ (the identity function $\Omega \to \Omega$), $\mu^{|\mathsf{X}}$ is a version of $\mu^{\mathsf{O}|\mathsf{X}}$. Note also that $\mathsf{Y} = \mathsf{Y} \circ \mathsf{O}$. Thus by Lemma 2.4.28, $\mathbb{K}$ is a version of $\mu^{\mathsf{Y}|\mathsf{X}}$.

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

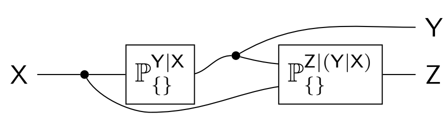
$$\mathbb{K}(\mathsf{Y} \bowtie A|x) = \delta_x(\mathsf{Y}^{-1}(A)) \tag{2.134}$$

$$= 0 \tag{2.135}$$

Thus $\mathbb{K}$ is valid. $\square$

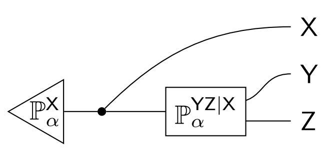
Theorem 2.4.33 (Higher order conditionals)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_C$ with conditional $\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}$. Then $\mathbb{P}_C^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ is a version of $\mathbb{P}_C^{\mathsf{Z}|\mathsf{YX}}$*

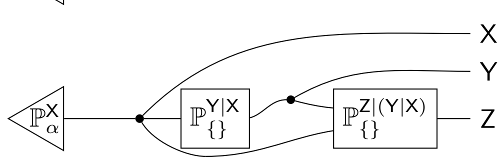
Proof. For arbitrary $\mathbb{P}_\alpha \in \mathbb{P}_C$



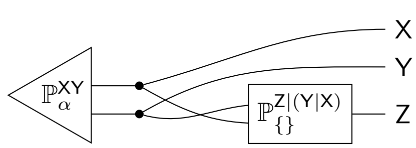
$$\mathbb{P}_\alpha^{\mathsf{YZ}|\mathsf{X}} = \tag{2.136}$$

$$\implies \mathbb{P}_\alpha^{\mathsf{XYZ}} = \tag{2.137}$$





$$= \tag{2.138}$$

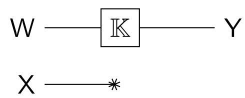


$$= \tag{2.139}$$

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

Theorem 2.4.34. *Given probability gap model $\mathbb{P}_C$, X, Y, Z such that $\mathbb{P}_C^{\mathsf{Z}|\mathsf{YX}}$ exists, $\mathbb{P}_C^{\mathsf{Z}|\mathsf{Y}}$ exists iff $\mathsf{Z} \perp\!\!\!\perp_{\mathbb{P}_C} \mathsf{X}|\mathsf{Y}$.*

Proof. If: If $\mathsf{Z} \perp\!\!\!\perp_{\mathbb{P}_C} \mathsf{X}|\mathsf{Y}$ then by Theorem 2.4.12, for each $\mathbb{P}_\alpha \in \mathbb{P}_C$ there exists $\mathbb{P}_\alpha^{\mathsf{Z}|\mathsf{Y}}$ such that

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


(2.140)

$\square$

Theorem 2.4.35 (Valid higher order conditionals)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_C$ with regular conditional $\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}$, Y discrete and Z standard measurable. Then there exists a valid regular $\mathbb{P}_C^{\mathsf{Z}|\mathsf{XY}}$.*

Proof. By Theorem 2.4.31, we have a higher order conditional $\mathbb{P}_C^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ which, by Theorem 2.4.33 is also a version of $\mathbb{P}_C^{\mathsf{Z}|\mathsf{XY}}$.

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

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

$$\forall A \in \mathcal{Z}, x \in X, y \in Y : \mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}(\{y\} \times A|x) = 0 \implies (\mathbb{Q}(A|x, y) = 0) \vee ((\mathsf{X}, \mathsf{Y}) \bowtie \{(x, y)\} = \emptyset) \tag{2.141}$$

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

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

Chapter 3

Models with choices and consequences

Probability sets, introduced in Chapter 2, will be used to model *decision problems*, which are problems that involve choices and consequences. In such problems, three things are given: a set of options, one of which must be chosen, a set of consequences and a means of judging which consequences are more desirable than others. Such a problem requires an understanding of how each choice corresponds to consequences, as far as this is able to be understood. The fundamental type of problem studied in this thesis is how to map choices to consequences.

In practice, causal inference is concerned with a wider variety of problems than this. A great deal of empirical causal analysis is concerned with problems a step removed from this: the purpose is to advise other decision makers on a course of action rather than to recommend an action directly. Nevertheless, a great deal of causal analysis is ultimately motivated by problems involving a choice among options, even if the analysis only addresses such problems indirectly. Section 3.1 briefly reviews the attitude of prominent theorists of causal inference towards decision problems. Subsequently, it presents the basic definition of a decision problem, and two different kinds of models that can be used to represent the relationship between choices and consequences.

The reasons we provide for using probability sets to model decision problems are not rigorous. The strongest motivation for this choice is *convention*: many varieties of decision theory induce probability set models, and Chapter 5 shows how many causal inference frameworks also induce probability set models. Some decision theories examined in this chapter justify their modelling choices by suggesting axioms for rational theories of decision under uncertainty. However, despite the various attempts at axiomatisation, the nature of theories of "rational choice" is contested – there is no clear standard among the the theories surveyed here, or developed elsewhere. This work is not trying to resolve this dispute, yet modelling choices must still be made. Section 3.2 provides an overview of four major decision theories along with their axiomatisations (where applicable). These are *Savage decision theory*, *Jeffrey decision theory* (or evidential decision theory), Lewis' *causal decision theory* and *statistical decision theory*.

Section 3.2 describes in particular detail the connections between *statistical decision theory* (Wald, 1950) and probability set models of decision problems. We are able to demonstrate a close connection between probability set models of decision problems and the classical statistical notion of *risk* of a decision rule, even though causal considerations are often not central to classical statistics. Secondly, the kind of probability set model – which we call a *see-do model* – shows up again in Chapter 4 where we consider the question of when a probability set model supports a certain notion of "the causal effect of a variable", and again in Chapter 5 where we consider the kinds of probability set models induced by other causal reasoning frameworks.

The formal definition of a variable in a probabilistic model is straightforward (Definition 2.2.13). However, in practice the definitions of variables often includes informal content that enables the interpretation of a probabilistic model. In the field of causal models, one is likely to come across many different "kinds" of variables: for example, observed variables, unobserved variables, counterfactual variables and causal variables all play important roles in various causal inference frameworks. However, there is no formal distinction between these different kinds of variables – Definition 2.2.13 applies to them all. Section 3.3 is an attempt to clarify an understanding of the informal role of variables as "pointing to the parts of the world that the model is about". In comparison to the wide variety of variable types encountered in the causal literature, it offers a very limited theory of the informal semantics of variables. In short, observed variables correspond to a measurement procedure (in a sense that will be made precise), and unobserved variables do not.

3.1 What is the point of causal inference?

Pearl and Mackenzie (2018) argues forcefully that causal reasoning frameworks should be understood by the questions that they answer. He also posits a "ladder" of types of causal question, where the nth level of question type also subsumes all lower levels. The question types are (Bareinboim et al., 2020):

1. *Associational*: informally, "questions about relationships and predictions"; formally defined as queries that can be answered by a single probability distribution

2. *Interventional*: informally, "questions about the consequences of interventions"; formally defined as queries that can be answered by a causal Bayesian network (CBN)

3. *Counterfactual*: informally, "questions concerning imagining alternate worlds"; formally defined as queries that can be answered by a structural causal model (SCM)

Given that counterfactual questions are suggested to be the most general kind of causal question, one might ask why this work focuses on questions of an interventional nature. There are two reasons for this: Firstly, a class of informal questions is being used to motivate the theory of causal inference with probability sets. I have much stronger intuitions about informal decision problems than informal counterfactual queries. This does not seem to be a purely personal taste: questions about how decision problems should be represented have been studied much more than similar questions regarding counterfactual queries. Secondly, problems that involve comparing different choices on the basis of their consequences are an important class of problems on their own. Even within the causal inference literature, "interventional" questions and interpretations are much more prominent than counterfactual questions. For example, Rubin (2005) points out that causal inference often informs a decision maker by providing "scientific knowledge", but does not make recommendations by itself. (Imbens and Rubin, 2015) introduces causal inference as the study of "outcomes of manipulations" and (Spirtes et al., 2000) highlights the universal relevance of understanding how to control certain outcomes, while further arguing that clarifying commonsense ideas of causation is also an important aim of causal inference. Hernán and Robins (2020) present causal knowledge as critical for assessing the consequences of actions.

Speculatively, counterfactual queries may be able to be interpreted as decision problems with fanciful options. Consider an informal decision problem and a counterfactual query addressing similar material:

- Decision problem: I want my headache to go away. If I take Aspirin, will it do so?
- Counterfactual query: I wish I didn't have headache. If I had taken the Aspirin, would I still have it?

If I haven't taken aspirin, then there's nothing I can actually choose to do to make it so that I had. However, if I imagine that I did have some option available that accomplished this, then the structure of the two questions seems rather similar. Both ask: if I take the option, what will the consequence be? Of course, it's hard to say what makes a correct answer to the second question, but this is a feature of counterfactual questions in general.

3.1.1 Modelling decision problems

People who need to make a decisions might (and often do) make them with no mathematical reasoning at all. However, this work is concerned with making decisions assisted by mathematical reasoning. In order to reason mathematically about a decision to be made, we assume that somehow, we have access to two sets:

1. There is a set of choices C that need to be compared
2. There is a set of consequences Ω along with a utility function $u : \Omega \to \mathbb{R}$

Given some means of relating between C and Ω, the order on Ω will induce some order on C. There are a great number of different ways that of relating elements of C to Ω. For example, a binary relation between the two sets will, given a total order on Ω, induce a preorder on C. However, in this work the assumption is made that the relevant kinds of relations are either

- A Markov kernel $C \rightarrowtriangle \Omega$
- A Markov kernel $C \times H \rightarrowtriangle \Omega$ for some set of hypotheses H

That is, for each choice $c \in C$ we have either a probability distribution in $\Delta(\Omega)$ or a set of probability distributions indexed by $h \in H$. Sections 3.2.5 and 3.2.5 discuss each choice in more detail. Where it is needed, we also assume that a utility function $\Omega \to \mathbb{R}$ is available and that choices are evaluated using the principle of expected utility.

Usually, someone confronted with a decision problem will not know for certain the consequences that arise from any given choice, and yet they may have some views about which consequences are more likely than others. Probability has a long and successful history of representing uncertain knowledge of this type. There are many works that aim to show that any method for representing uncertain knowledge that adheres to certain principles must be a probability distribution de Finetti ([1937] 1992); Horvitz et al. (1986), along with criticism of these principles Halpern (1999). A notable alternative to representing uncertainty with a single probability distribution represents uncertainty with a set of probability distributions, which is a type of *vague probability* model (Walley, 1991).

More relevant to the question of modelling decision problems are a number of works that establish conditions under which "desirability" or "preference" relations over sets of choices or propositions must be represented by a probability distribution along with a utility function. These works are surveyed in Section 3.2. Ultimately, however, the question of whether probability is the right choice to represent uncertain knowledge in decision models is not a key focus of this work. It is a conventional choice, and one that is accepted here.

3.1.2 Formal definitions

We suppose that we are, at the outset, given a few basic ingredients: a set of choices C, a set of consequences Ω and a utility function $u : \Omega \to \mathbb{R}$. We call these ingredients a "decision problem".

Definition 3.1.1 (Decision problem)**.** A decision problem is a triple (C, Ω, u) consisting of a measurable set $(C, \mathcal{C})$ of choices, $(\Omega, \mathcal{F})$ consequences and a utility function $u : \Omega \to \mathbb{R}$.

Our task is to find a *model* that relates C to Ω. We assume two forms of model – a *sharp model* associates each choice with a unique probability distribution, and a *vague model* associates each choice with a set of probability distributions.

Definition 3.1.2 (Choices only model)**.** Given a decision problem (C, Ω, u), a *choices only model* is a function $C \rightarrowtriangle \Omega$.

Definition 3.1.3 (Choices and hypotheses model)**.** Given a decision problem (C, Ω, u), a model with *choices and hypotheses* is a function $C \times H \rightarrowtriangle \Omega$ for some hypothesis set H.

Both types of models induce probability sets.

Definition 3.1.4 (Induced probability set)**.** Given a decision problem (C, Ω, u) and a model $\mathbb{P}_\cdot : C \times H \rightarrowtriangle \Omega$, the induced probability set is $\mathbb{P}_{C\times H} := \{\mathbb{P}_\alpha | \alpha \in C \times H\}$.

3.2 Representation theorems for decision problems

We assume decision models are probabilistic functions $C \to \Delta(\Omega)$ for some sample space $(\Omega, \mathcal{F})$ of "consequences". Probability distributions, and the principle of expected utility in particular, are common choices for evaluation under uncertainty. Representation theorems offer a more formal justification for this choice; they propose a collection of axioms regulating the sets of evaluations we want some decision evaluation model to admit, and then show that this model can be represented with (for example) a probability distribution along with a utility function. The desirability of some of the axioms in these theorems is not obvious.

Here we review the representation theorems of Savage (1954) and Jeffrey (1965). We establish that both imply that choices are compared using a probabilistic function $C \to \Delta(\Omega)$ for a suitable selection of C and $(\Omega, \mathcal{F})$, along with a "desirability" function which differs in type between the two theorems.

Lewis' *causal decision theory* is also briefly reviewed. While the particular considerations that motivated this theory are not examined, this theory introduces *dependency hypotheses*, which play a key role in the rest of this work.

The following discussion will often make reference to *complete preference relations*. A complete preference relation is a relation $\succ, \prec, \sim$ on a set A such that for any a, b, c in A we have:

- Exactly one of $a \succ b$, $a \prec b$, $a \sim b$ holds
- $(a \succ b) \iff (b \prec a)$
- $a \succ b$ and $b \succ c$ implies $a \succ c$

In short, it is a total order without antisymmetry (a and b can be equally preferred even if they are not in fact equal).

This definition is meant to correspond to the common sense idea of having preferences over some set of things, where $\succ$ can be read as "strictly better than", $\prec$ read as "strictly worse than" and $\sim$ read as "as good as". Given any two things from the set, I can say which one I prefer, or if I prefer neither (and all of these are mutually exclusive). If I prefer a to a' then I think a' is worse than a. Furthermore, if I prefer a to a' and a' to a'' then I prefer a to a''.

Define $a \preceq b$ to mean $a \prec b$ or $a \sim b$.

3.2.1 von Neumann-Morgenstern utility

Von Neumann and Morgenstern (1944) proved that when the *vNM axioms* hold (not defined here; see the original reference or Steele and Stefánsson (2020)), an agent's preferences between "lotteries" (probability distributions in $\Delta(\Omega)$ for some $(\Omega, \mathcal{F})$) can be represented as the comparison of the expected value under each lottery of a utility function u unique up to affine transformation. That is, for lotteries $\mathbb{P}_\alpha$ and $\mathbb{P}_{\alpha'}$, there exists some $u : \Omega \to \mathbb{R}$ unique up to affine transformation such that $\mathbb{E}_{\mathbb{P}_\alpha}[u] > \mathbb{E}_{\mathbb{P}_{\alpha'}}[u]$ if and only if $\mathbb{P}_\alpha \succ \mathbb{P}_{\alpha'}$.

In vNM theory, the set of lotteries is is the set of all probability measures on $(\Omega, \mathcal{F})$. Thus von Neumann-Morgenstern theorem gives conditions under which preferences *over distributions of consequences* can be represented using expected utility. It is silent on the question of whether each choice should be mapped to a unique probability distribution over consequences.

3.2.2 Savage decision theory

Savage's decision theory establishes conditions under which, given *acts* C, *consequences* Ω and *states* $(S, \mathcal{S})$ (which are "possible mappings from acts to consequences"), the preference relation over acts can be represented with a probability distribution over states and a utility function $\Omega \to \mathbb{R}$. This is much closer to the subject of this work than the theorem of von Neumann and Morgernstern.

Definition 3.2.1 (Elements of a Savage decision problem)**.** A *Savage decision problem* features a measurable set of states $(S, \mathcal{S})$, a set of consequences Ω and a set of acts C such that $|C| = \Omega^S$ and an evaluation function $T : S \times C \to F$ such that for any $f : S \to \Omega$ there exists $c \in C$ such that $T(\cdot, c) = f$.

Theorem 3.2.2. *Given any Savage decision problem (S, Ω, C, T) with a preference relation $(\prec, \sim)$ on C that satisfies the* Savage axioms *3.2.2, there exists a unique probability distribution $\mu \in \Delta(\mathcal{S})$ and a utility $u : \Omega \to \mathbb{R}$ unique up to affine transformation such that*

$$\alpha \preceq \alpha' \iff \int_S u(T(s, \alpha))\mu(\mathrm{d}s) \leq \int_S u(T(s, \alpha'))\mu(\mathrm{d}s) \qquad \forall \alpha, \alpha' \in C \tag{3.1}$$

Proof. Savage (1954) $\square$

If we equip consequences with a measures $(\Omega, \mathcal{F})$, Savage's setup implies the existence of a unique probabilistic function $C \to \Delta(\Omega)$ representing the "probabilistic consequences" of each choice.

Theorem 3.2.3. *Given any Savage decision problem (S, Ω, C, T) with a preference relation $(\prec, \sim)$ on C that satisfies the* Savage axioms*, and a σ-algebra $\mathcal{F}$ on Ω such that T is measurable, there is a probabilistic function $\mathbb{P}_{\cdot} : C \to \Delta(\Omega)$ and a utility $u : \Omega \to \mathbb{R}$ unique up to affine transformation such that*

$$\alpha \preceq \alpha' \iff \int_\Omega u(f)\mathbb{P}_\alpha(\mathrm{d}f) \leq \int_\Omega u(f)\mathbb{P}_{\alpha'}(\mathrm{d}f) \qquad \forall \alpha, \alpha' \in C \tag{3.2}$$

Proof. Define $\mathbb{P}_{\cdot} : C \to \Delta(\Omega)$ by

$$\mathbb{P}_\alpha(A) := \mu(T_\alpha^{-1}(A)) \qquad \forall A \in \mathcal{F} \tag{3.3}$$

where $\mathsf{T}_\alpha : S \to F$ is the function $s \mapsto T(s, \alpha)$. $\mathbb{P}_\alpha$ is the pushforward of T_α under μ.

Then

$$\int_\Omega u(f)\mathbb{P}_\alpha(\mathrm{d}f) = \int_S u \circ T_\alpha(s)\mu(\mathrm{d}s) \tag{3.4}$$

$$= \int_S u(T(s, \alpha))\mu(\mathrm{d}s) \tag{3.5}$$

□

Savage axioms

Careful analysis of Savage's theorem is outside the scope of this work, but given the relevant of Savage's representation theorem we will reproduce the axioms from Savage (1954) with a small amount of commentary. Keep in mind that Savage's theorem establishes that the following are sufficient for representation with a probability set, not necessary, and furthermore the probability set representation of preferences satisfying these axioms is unique.

Given acts C, states $(S, \mathcal{S})$ and consequences F and a map $T : S \times C \to F$, let all greek letters α, β etc. be elements of C. Savage's axioms are:

P1: There is a complete preference relation $\preceq$ on C

> D1: $\alpha \preceq \beta$ given $B \in \mathcal{S}$ if and only if $\alpha' \preceq \beta'$ for every α' and β' such that $T(\alpha, s) = T(\alpha', s)$ for $s \in B$ and $T(\alpha', r) = T(\beta', r)$ for $r \notin B$, and $\beta' \preceq \alpha'$ either for every such pair or for none.

P2: For every α, β and $B \in \mathcal{S}$, $\alpha \preceq \beta$ given B or $\beta \preceq \alpha$ given B

> D2: for $q, q' \in F$, $q \preceq q'$ if and only if $\alpha \preceq \alpha'$ where $T(\alpha, s) = q$ and $T(\alpha', s) = q'$ for all $s \in S$
>
> D2: $B \in \mathcal{S}$ is null if and only if $\alpha \preceq \beta$ given B for every $\alpha, \beta \in C$

P3: If $T(\alpha, s) = q$ and $T(\alpha', s) = q'$ for every $s \in B$, $B \in \mathcal{S}$ non-null, then $\alpha \preceq \alpha'$ given B if and only if $q \preceq q'$

D4: For $A, B \in \mathcal{S}$, $A \leqslant B$ if and only if $\alpha_A \preceq \alpha_B$ or $q \preceq q'$ for all $\alpha_A, \alpha_B \in C$, $q, q' \in F$ such that $T(\alpha_A, s) = q$ for $s \in A$, $T(\alpha_A, s') = q'$ for $s' \notin A$, $T(\alpha_B, s) = q$ for $s \in B$, $T(\alpha_B, s') = q'$ for $s' \notin B$. Read $\leqslant$ as "is less probable than"

P4: For every $A, B \in \mathcal{S}$, $A \leqslant B$ or $B \leqslant A$

P5: For some α, β, $\alpha \prec \beta$

P6: Suppose $\alpha \not\preceq \beta$. Then for every γ there is a finite partition of S such that if α' agrees with α and β' agrees with β except on some element B of the partition, α' and β' being equal to γ on B, then $\alpha \not\preceq \beta'$ and $\alpha' \not\preceq \beta$

D5: $\alpha \preceq q$ for $q \in F$ given B if and only if $\alpha \preceq \beta$ given B where $T(\beta, s) = q$ for all $s \in S$

P7: If $\alpha \preceq T(\beta, s)$ given B for every $s \in B$, then $\alpha \preceq \beta$ given B

P7': The proposition given by inverting every expression in D5 and P7

Our initial view of decision problems was that the consequences Ω are a set of things we know how to rank and choices C are the things we want to rank. This is not exactly Savage's setup – he assumes a preference relation ranking "acts" C to begin with. Furthermore, Savage also introduces a set of states S and assumes that the set of acts corresponds to the set of all function $S \to \Omega$. Many decision problems might be able to be extended with states and the set of acts enriched so as to satisfy these requirements, but it is not obvious that this is always possible.

D1 formalises the idea of one act α being not preferred to another β given the knowledge that the true state lies in the set B (in short: "given B" or "conditional on B"). P2 is sometimes called the "sure thing principle", as it implies the following: for any α, β if α is better than β on some states and no worse on any other, then $\alpha \succ \beta$. In Savage's model, the "likelihood" that of any state cannot depend on the act chosen.

D4 + P4 defines the "probability preorder" $\leqslant$ on $(S, \mathcal{S})$ and assumes it is complete.

P5 is the requirement that the preference relation is non-trivial; not everything is equally desirable. This doesn't seem like it should be a practical requirement to me; we might hope that a model can distinguish between some of our options, but that doesn't mean we should assume it can. Savage claims that this requirement is "innocuous" because any exception must be trivial, but I'm not sure I agree.

P6 is a requirement of continuity; for any $\alpha \preceq \beta$, we can divide S finely enough to squeeze a "small slice" of any third outcome γ into the gap between the two.

P7 in combination with the other axioms forces preferences to be bounded.

3.2.3 Jeffrey's decision theory

Jeffrey's decision theory is an alternative to Savage's that starts from a different set of assumptions. One of the key differences is in what is assumed at the outset: where Savage assumes a set of states S, acts C and consequences Ω, Jeffrey's theory only considers a single space $\underline{\mathcal{F}}$, which is a complete atomless boolean algebra. Elements of $\underline{\mathcal{F}}$ are said to be propositions, although the structure of $\underline{\mathcal{F}}$ means we can't understand it as, for example, a set of propositions regarding the result of a particular measurement procedure (Section 3.3). The theory is set out in Jeffrey (1965), and the key representation theorem proved in Bolker (1966).

Recall that our fundamental problem is relating a set C of things we can choose to a set F of things we can compare. Jeffrey's theory uses a different strategy to accomplish this than Savages'; where identifies a set of acts C with all functions $S \to F$ and proposes axioms that constrain a preference relation on C, Jeffrey assumes that choices are elements of the algebra $\underline{\mathcal{F}}$, along with propositions that do not correspond to choices. Jeffrey's axioms pertain to a preference relation on $\underline{\mathcal{F}}$. The ultimate result is, for our purposes, very similar.

Theorem 3.2.4. *Suppose there is a complete atomless Boolean algebra $\underline{\mathcal{F}}$ with a preference relation $\preceq$. If $\preceq$ satisfies the* Bolker axioms *(Section 3.2.3) then there exists a desirability function $des : \underline{\mathcal{F}} \to \mathbb{R}$ and a probability distribution $\mu \in \Delta(\underline{\mathcal{F}})$ such that for $A, B \in \underline{\mathcal{F}}$ and finite partition $D_1, ..., D_n \in \underline{\mathcal{F}}$:*

$$(A \preceq B) \iff \sum_i^n des(D_i)\mu(D_i|A) \leq \sum_i^n des(D_i)\mu(D_i|B) \tag{3.6}$$

where $\mu(D_i|A) := \frac{\mu(A \cap D_i)}{\mu(A)}$ for $\mu(A) > 0$, undefined otherwise.

Proof. Bolker (1966)) □

As mentioned, in Jeffrey's theory the *choices* C are a subset of $\underline{\mathcal{F}}$. Thus we can deduce from a Jeffrey model a function $C \to \Delta(\underline{\mathcal{F}})$ that "represents the consequences of choices" in the sense of Theorem 3.2.5.

Theorem 3.2.5. *Suppose there is a complete atomless Boolean algebra $\underline{\mathcal{F}}$ with a preference relation $\preceq$ that satisfies the Bolker axioms, and a set of choices C over which a preference relation is sought with $\mu(\alpha) > 0$ for all $\alpha \in C$. Then there is a function $\mathbb{P}_\cdot : C \to \Delta(\underline{\mathcal{F}})$ such that for any $\alpha, \alpha' \in C$ and finite partition $D_1, ..., D_n \in \underline{\mathcal{F}}$:*

$$\alpha \preceq \alpha' \iff \sum_i^n des(D_i)\mathbb{P}_\alpha(D_i) \leq \sum_i^n des(D_i)\mathbb{P}_{\alpha'}(D_i) \tag{3.7}$$

Where μ and des *are as in Theorem 3.2.4*

Proof. Define $\mathbb{P}_\cdot$ by $\alpha \mapsto \mu(\cdot|\alpha)$. Then Equation 3.7 follows from Equation 3.6. □

Bolker axioms

$\underline{\mathcal{F}}$ a complete, atomless Boolean algebra with the impossible proposition. An example of such a set is constructed from the set of Lebesgue measurable sets on $[0,1]$ identifying any two sets that differ by a set of measure zero identified Bolker (1967). This is not a σ-algebra.

A1: $\preceq$ is a complete preference relation

B2: $\underline{\mathcal{F}}$ is a complete, atomless Boolean algebra with the impossible proposition removed

C3: For $A, B \in \underline{\mathcal{F}}$, if $A \cap B = \emptyset$, then

a) If $A \succ B$ then $A \succ A \cup B \succ B$

b) If $A \sim B$ then $A \sim A \cup B \sim B$

D4: Given $A \cap B = \emptyset$ and $A \sim B$, if $A \cup G \sim B \cup G$ for some G where $A \cap G = B \cap G = \emptyset$ and $G \not\sim A$, then $A \cup G \sim B \cup G$ for every such G

D1: The supremum (infimum) of a subset $W \subset \underline{\mathcal{F}}$ is a set G (D) such that for all $A \in W$, $G \subset A$ $(A \subset D)$, and for any E that also has this property, $G \subset E$ $(E \subset D)$

E5: Given $W := \{W_i\}_{i \in M \subset \mathbb{N}}$ with $i < j \implies W_j \subset W_i$ and $W \subset \underline{\mathcal{F}}$ with supremum G (infimum D), whenever $A \prec G \prec B$ $(A \prec D \prec B)$ then there exists some $k \in M$ such that $i \geq k$ $(i \leq k)$ implies $A \prec W_i \prec B$.

Like Savage's theory, A1 requires the preference relation to be complete.

A3 is the assumption that the desirability of disjunctions of events lies between the desirability of each event; it is sometimes called "averaging". It notably rules out the following: if $A \succ B$ we cannot have $A \cup B \sim A$. In the Jeffrey-Bolker theory, propositions all have positive probabilities.

A4 allows a probability order to be defined on $\underline{\mathcal{F}}$. The conditions $A \cap B = \emptyset$, $A \sim B$, $A \cup G \sim B \cup G$ for some G where $A \cap G = B \cap G = \emptyset$ and $G \not\sim A$ can be seen as a test for A and B being "equally probable". A4 requires that if A and B are rated as equally probable by one such test, then they are rated as equally probable by all such tests.

A5 is an axiom of continuity.

3.2.4 Causal decision theory

Causal decision theory was developed after both Jeffrey's and Savage's theory. A number of authors Lewis (1981); Skyrms (1982) felt that Jeffrey's theory erred by treating the consequences of a choice as an "ordinary conditional probability". Lewis (1981) suggested that causal decision theory can be used to evaluate choices when we are given a set Ω of consequences over which preferences are known, a set C of choices and a set H of dependency hypotheses (the letters have been changed to match usage in this work; in the original the consequences were called S, the choices A and the dependency hypotheses H). Choices are then evaluated according to the causal decision rule. We have taken the liberty to state Lewis' rule in the language of the present work.

Definition 3.2.6 (Causal decision rule)**.** Given a set C of choices, sample space $(\Omega, \mathcal{F})$, variables $\mathsf{H} : \Omega \to H$ (the *dependency hypothesis*) and $\mathsf{S} : \Omega \to S$ (the *consequence*) and a utility $u : \Omega \to \mathbb{R}$, the *causal utility* of a choice $\alpha \in C$ is given by

$$U(\alpha) := \int_S \int_H u(s) \mathbb{P}_\alpha^{\mathsf{S}|\mathsf{H}}(\mathrm{d}s|h) \mathbb{P}_C^{\mathsf{H}}(\mathrm{d}h) \tag{3.8}$$

For some probabilistic function $\mathbb{P}_\cdot : C \to \Delta(\Omega)$.

The reasons why Lewis wanted to introduce dependency hypothesis and modify Jeffrey's rule to Equation 3.8 are controversial and do not come up in this work. However, causal decision theory is still relevant to this work in two ways: firstly, once again is a probabilistic function $\mathbb{P}_\cdot : C \to \Delta(\Omega)$. Secondly, causal decision theory introduces the notion of the dependency hypothesis H. The dependency hypothesis is similar to the state in Savage's theory, however Lewis does not require a deterministic map from dependency hypotheses to consequences, nor does he require a choice to correspond to every possible function from dependency hypotheses to states.

Dependency hypotheses are quite an important idea in causal reasoning. Together Lewis' decision rule connect the theory of probability sets with *statistical decision theory*, as Section 3.2.5 will show. Chapter 4 goes into considerable detail concerning the question of when probability sets support certain types of dependency hypothesis. While they are typically not explicitly represented in common frameworks for causal inference, Chapter 5 discusses how dependency hypotheses are often implicit in these approaches, and shows how they can be made explicit.

3.2.5 Statistical decision theory

Statistical decision theory (SDT), created by Wald (1950), predates all of the decision theories discussed above. Savage's theory appears to have developed in part to explain some features of SDT Savage (1951), and Jeffrey's theory and subsequent causal decision theories were in turn influenced by Savage's decision theory. While the later decision theories were concerned with articulating why their theory fit the role of a theory for rational decision under uncertainty, Wald focused much more on the mathematical formalism and solutions to statistical problems. Statistical decision theory introduced many fundamental ideas that have since entered the "water supply" of machine learning theory, such as *decision rules* and *risk* as a measure of the quality of a decision rule.

In contrast to the later decision theories, SDT has no explicit representation of the "consequences" of a decision. Rather, it is assumed that a loss function is given that maps decisions and hypotheses directly to a loss, which is a kind of desirability score similar to a utility (although it is minimised rather than maximised).

Definition 3.2.7 (Statistical decision problem)**.** A statistical decision problem (SDP) is a tuple $(X, H, D, l, \mathbb{P}.)$ where $(X, \mathcal{X})$ is a set of outcomes, $(H, \mathcal{H})$ is a set of hypotheses, $(D, \mathcal{D})$ is a set of decisions, $l : D \times H \to \mathbb{R}$ is a loss function and $\mathbb{P}. : H \rightarrow X$ is a Markov kernel from hypotheses to to outcomes.

Statistical decision theory is concerned with the selection of *decision rules*, rather than the selection of decisions directly. A decision rule maps observations to decisions, and may be deterministic or stochastic.

Definition 3.2.8 (Decision rule)**.** Given a statistical decision problem $(X, H, D, l, \mathbb{P}.)$, a decision rule is a Markov kernel $\mathbb{D}_\alpha : \Omega \rightarrow D$.

Because decision rules in SDT play the role of what we call *choices*, we denote the set of all available decision rules by C. A further feature of SDT that is unlike the later decision theories is that SDT does not offer a single rule for assessing the desirability of any choice in C. Instead, it offers a definition of the risk, which assesses the desirability of a choice *relative to a particular hypothesis*. The risk function completely characterises the problem of choosing a decision function. Two different rules are for turning this "intermediate assessment" into a final assessment of the available choices - Bayes optimality and minimax optimality. Bayes optimality reuquires a prior over hypotheses, while minimax optimality does not.

Definition 3.2.9 (SDP Risk)**.** Given a statistical decision problem $(X, H, D, l, \mathbb{P}.)$ and decision functions C, the *risk* functional $R : C \times H \to \mathbb{R}$ is defined by

$$R(\mathbb{D}_\alpha, h) := \int_X \int_D l(d, h)\mathbb{D}_\alpha(\mathrm{d}d|f)\mathbb{P}_h(\mathrm{d}f) \tag{3.9}$$

It is possible to find risk functions in problems that aren't SDPs. The definitions of Bayes and Minimax optimality still apply to risk functions obtained on other manners. Thus Bayes optimality and minimax optimality are defined in terms of risk functions in general, not SDP risk functions.

Definition 3.2.10 (Bayes risk)**.** Given decision functions C, hypotheses $(H, \mathcal{H})$, risk $R : C \times H \to \mathbb{R}$ and prior $\mu \in \Delta(H)$, the *μ-Bayes risk* is

$$R_\mu(\mathbb{D}_\alpha) := \int_H R(\mathbb{D}_\alpha, h)\mu(\mathrm{d}h) \tag{3.10}$$

Definition 3.2.11 (Bayes optimal)**.** Given decision functions C, hypotheses $(H, \mathcal{H})$, risk $R : C \times H \to \mathbb{R}$ and prior $\mu \in \Delta(H)$, $\alpha \in C$ is μ-Bayes optimal if

$$R_\mu(\mathbb{D}_\alpha) = \inf_{\alpha' \in C} R_\mu(\mathbb{D}_{\alpha'}) \tag{3.11}$$

Definition 3.2.12 (Minimax optimal)**.** Given decision functions C, hypotheses $(H, \mathcal{H})$, risk $R : C \times H \to \mathbb{R}$, a *minimax decision function* is any decision function $\mathbb{D}_\alpha$ satisfying

$$\sup_{h\in H} R(\mathbb{D}_\alpha, h) = \inf_{\alpha' inC} \sup_{h\in H} R(\mathbb{D}_{\alpha'}, h) \tag{3.12}$$

From consequences to statistical decision problems

Statistical decision theory ignores the notion of general consequences of choices; the only "consequence" in the theory is the loss incurred by a particular decision under a particular hypothesis. The kinds of probability set models studied here probabilistically map decisions to consequences, and the set of consequences is understood to have a utility function to allow for assessment of the desirability of different choices via the principle of expected utility. Not every probability set model induces a statistical decision problem in this manner. A family of models that does are what we call *conditionally independent see-do models*. These models feature observations (the "see" part) along with decisions and consequences (the "do" part), and the observations come "before" the decisions (hence see-do). Examples of this type of model will be encountered again in Chapters 4 and 5. Furthermore, there is a hypothesis such that consequences are assumed to be independent of observations conditional on the decision and the hypothesis. This is why they are qualified as "conditionally independent" see-do models.

Definition 3.2.13 (See-do model)**.** A probability set model of a statistical decision problem, or a *see-do model* for short, is a tuple $(\mathbb{P}_{C\times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D})$ where $\mathbb{P}_{C\times H}$ is a probability set indexed by elements of $C \times H$ on $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$ are the observations, $\mathsf{Y} : \Omega \to Y$ are the consequences and $\mathsf{D} : \Omega \to D$ are the decisions. $\mathbb{P}_{C\times H}$ must observe the following conditional independences:

$$\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_{C\times H}} \mathsf{C}|\mathsf{H} \tag{3.13}$$

$$\mathsf{D} \perp\!\!\!\perp^e_{\mathbb{P}_{C\times H}} \mathsf{H}|\mathsf{C} \tag{3.14}$$

where $\mathsf{C} : C \times H \to C$ and $\mathsf{H} : C \times H \to H$ are the respective projections (see Definition 2.4.16 for the definition of extended conditional independence).

Definition 3.2.14 (Conditionally independent see-do model)**.** A conditionally independent see-do model is a see do model $(\mathbb{P}_{C\times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D})$ where the following additional conditional independence holds:

$$\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_{C\times H}} (\mathsf{X}, \mathsf{C})|(\mathsf{D}, \mathsf{H}) \tag{3.15}$$

We assume that a utility function is available depending on the consequence Y only, and identify the loss with the negative expected utility, conditional on a particular decision and hypothesis.

Definition 3.2.15 (Induced loss)**.** Given a see-do model $(\mathbb{P}_{C\times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D})$ and a utility $u : Y \to \mathbb{R}$, the induced loss $l : D \times H \to \mathbb{R}$ is defined as

$$l(d, h) := -\int_Y u(y)\mathbb{P}^{\mathsf{Y}|\mathsf{D}}_{C\times\{h\}}(\mathrm{d}y|d) \tag{3.16}$$

where the uniform conditional $\mathbb{P}^{\mathsf{Y}|\mathsf{D}}_{C\times\{h\}}$'s existence is guaranteed by $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_{C\times H}} (\mathsf{X}, \mathsf{C})|(\mathsf{D}, \mathsf{H})$.

A see-do model induces a set of decision functions: for each $\alpha \in C$, there is an associated probability distribution $\mathbb{P}^{\mathsf{D}|\mathsf{X}}_\alpha$. Using the above definition of loss, the expected loss of a decision function in a conditionally independent see-do model induces a risk function identical to the SDP risk.

Theorem 3.2.16 (Induced SDP risk)**.** *Given a conditionally independent see-do model* $(\mathbb{P}_{C\times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D})$ *along with a utility* $u : Y \to \mathbb{R}$, *the expected utility for each choice* $\alpha \in C$ *and hypothesis* $h \in H$ *is equal to the negative SDP risk of the associated decision rule* $\mathbb{P}^{\mathsf{D}|\mathsf{X}}_\alpha$ *and hypothesis* h.

$$\mathbb{P}^{\mathsf{Y}}_{\alpha,h} u = -R(\mathbb{P}^{\mathsf{D}|\mathsf{X}}_{\{\alpha\}\times H}, h) \tag{3.17}$$

Proof. The expected utility given α and h is

$$\begin{aligned}
\int_Y u(y)\mathbb{P}^{\mathsf{Y}}_{\alpha,h}(\mathrm{d}y) &= \int_Y\int_D\int_X u(y)\mathbb{P}^{\mathsf{Y}|\mathsf{DX}}_{\alpha,h}(\mathrm{d}y|d,x)\mathbb{P}^{\mathsf{D}|\mathsf{X}}_{\alpha,h}(\mathrm{d}d|x)\mathbb{P}^{\mathsf{X}}_{\alpha,h}(\mathrm{d}x) &(3.18)\\
&= \int_X\int_D\int_Y u(y)\mathbb{P}^{\mathsf{Y}|\mathsf{D}}_{\alpha,h}(\mathrm{d}y|d)\mathbb{P}^{\mathsf{D}|\mathsf{X}}_{\alpha,h}(\mathrm{d}d|x)\mathbb{P}^{\mathsf{X}}_{\alpha,h}(\mathrm{d}x) &(3.19)\\
&= \int_X\int_D\int_Y u(y)\mathbb{P}^{\mathsf{Y}|\mathsf{D}}_{C\times\{h\}}(\mathrm{d}y|d)\mathbb{P}^{\mathsf{D}|\mathsf{X}}_{\{\alpha\}\times H}(\mathrm{d}d|x)\mathbb{P}^{\mathsf{X}}_{C\times\{h\}}(\mathrm{d}x) &(3.20)\\
&= -\int_D\int_X l(d,h)\mathbb{P}^{\mathsf{D}|\mathsf{X}}_{\{\alpha\}\times H}(\mathrm{d}d|x)\mathbb{P}^{\mathsf{X}}_{C\times\{h\}}(\mathrm{d}x) &(3.21)\\
&= -R(\mathbb{P}^{\mathsf{D}|\mathsf{X}}_{\{\alpha\}\times H}, h) &(3.22)
\end{aligned}$$

where Equation 3.19 follows from $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_{C\times H}} (\mathsf{X}, \mathsf{C})|(\mathsf{D}, \mathsf{H})$, the uniform conditional $\mathbb{P}^{\mathsf{D}|\mathsf{X}}_{\{\alpha\}\times H}$ exists due to $\mathsf{D} \perp\!\!\!\perp^e_{\mathbb{P}_{C\times H}} \mathsf{H}|\mathsf{C}$ and the uniform conditional $\mathbb{P}^{\mathsf{X}}_{C\times\{h\}}$ exists due to $\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_{C\times H}} \mathsf{C}|\mathsf{H}$. □

Theorem 3.2.16 does *not* hold for general see-do models. General see-do models allow for the utility to depend on X even after conditioning on D and H, while the form of the loss function in SDT forces no direct dependence on observations. The generic "see-do risk" (Definition 3.2.17) provides a notion of risk for the more general case, while Theorem 3.2.16 shows it reduces to SDP risk in the case of conditionally independent see-do models with a utility that depends only on the consequences Y.

Definition 3.2.17 (See-do risk)**.** Given a see-do model $(\mathbb{P}_{C\times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D})$ along with a utility $u : X \times Y \to \mathbb{R}$, the *see-do risk* $R : C \times H \to \mathbb{R}$ is given by

$$R(\alpha, h) := -\mathbb{P}^{\mathsf{XY}}_{\alpha,h} u \qquad\qquad \forall \alpha \in C, h \in H \tag{3.23}$$

Section 3.1.1 noted that two types of probability set model are considered: probability sets $\mathbb{P}_C$ indexed by choices alone, and probability sets $\mathbb{P}_{C\times H}$ jointly indexed by choices and hypotheses. See-do models are an instance of the second kind, jointly indexed by choices and hypotheses. Bayesian see-do models are of the former type, indexed by choices alone. A see-do model $(\mathbb{P}_{C\times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D})$ and a prior over hypotheses $\mu \in \Delta(H)$ can by combined to form a Bayesian see-do model, and under the right conditions the risk of the Bayesian model reduces to the Bayes risk of the original see-do model.

Definition 3.2.18 (Bayesian see-do model)**.** A Bayesian see-do model is a tuple $(\mathbb{P}_C, \mathsf{X}, \mathsf{Y}, \mathsf{D}, \mathsf{H})$ where $\mathbb{P}_C$ is a probability set on $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$ are the observations, $\mathsf{Y} : \Omega \to Y$ are the consequences, $\mathsf{D} : \Omega \to D$ are the decisions and $\mathsf{H} : \Omega \to H$ is the hypothesis. $\mathbb{P}_C$ must observe the following conditional independences:

$$\mathsf{X} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{C}|\mathsf{H} \tag{3.24}$$
$$\mathsf{D} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{H}|\mathsf{C} \tag{3.25}$$
$$\mathsf{H} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{C} \tag{3.26}$$

Definition 3.2.19 (Induced Bayesian see-do model)**.** Given a see-do model $(\mathbb{P}_{C\times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D}, \mathsf{H})$ on $(\Omega, \mathcal{F})$ and a prior $\mu \in \Delta(H)$, the induced Bayesian see-do model $\mathbb{P}_C$ on $(\Omega \times H, \mathcal{F} \otimes \mathcal{H})$ is

$$\mathbb{P}_C(A) = \int_{\mathsf{H}^{-1}(A)} \mathbb{P}_{C\times\{h\}}(\Pi_\Omega^{-1}(A))\mu(\mathrm{d}h) \qquad \forall A \in \mathcal{F} \otimes \mathcal{H} \tag{3.27}$$

Where $\Pi_\Omega : \Omega \times H \to \Omega$ is the projection onto Ω.

Theorem 3.2.20 (Induced SDP Bayes risk)**.** *Given a conditionally independent see-do model* $(\mathbb{P}_C, \mathsf{X}, \mathsf{Y}, \mathsf{D}, \mathsf{H})$ *along with a utility* $u : Y \to \mathbb{R}$ *and a prior* $\mu \in \Delta(H)$*, the expected utility for each choice* $\alpha \in C$ *under the induced Bayesian see-do model is equal to the negative* μ*-Bayes risk of that decision rule.*

Proof. First, note that $h \mapsto \mathbb{P}^{\mathsf{Y}|\mathsf{XD}}_{C\times\{h\}}$ is a version of $\mathbb{P}^{\mathsf{Y}|\mathsf{XD}}_C$ and hence $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{X}, \mathsf{C})|(\mathsf{H}, \mathsf{D})$, a property it inherits from the underlying see-do model.

Also, note that $\mathbb{P}^{\mathsf{H}}_C = \mu$, by construction.

The expected utility of $\alpha \in C$ is

$$
\begin{aligned}
\mathbb{P}^{\mathsf{Y}}_\alpha u &= \int_Y u(y)\mathbb{P}^{\mathsf{Y}}_\alpha(\mathrm{d}y) && (3.28)\\
&= \int_Y \int_D \int_X \int_H u(y)\mathbb{P}^{\mathsf{Y}|\mathsf{DXH}}_\alpha(\mathrm{d}y|d,x,h)\mathbb{P}^{\mathsf{D}|\mathsf{XH}}_\alpha(\mathrm{d}d|x,h)\mathbb{P}^{\mathsf{X}|\mathsf{H}}_\alpha(\mathrm{d}x|h)\mathbb{P}^{\mathsf{H}}_\alpha(\mathrm{d}h) && (3.29)\\
&= \int_X \int_D \int_Y \int_H u(y)\mathbb{P}^{\mathsf{Y}|\mathsf{DH}}_\alpha(\mathrm{d}y|d,h)\mathbb{P}^{\mathsf{D}|\mathsf{X}}_\alpha(\mathrm{d}d|x)\mathbb{P}^{\mathsf{X}|\mathsf{H}}_\alpha(\mathrm{d}x|h)\mathbb{P}^{\mathsf{H}}_\alpha(\mathrm{d}h) && (3.30)\\
&= \int_X \int_D \int_Y \int_H u(y)\mathbb{P}^{\mathsf{Y}|\mathsf{DH}}_C(\mathrm{d}y|d,h)\mathbb{P}^{\mathsf{D}|\mathsf{X}}_\alpha(\mathrm{d}d|x)\mathbb{P}^{\mathsf{X}|\mathsf{H}}_C(\mathrm{d}x|h)\mu(\mathrm{d}h) && (3.31)\\
&= -\int_D \int_X \int_H l(d,h)\mathbb{P}^{\mathsf{D}|\mathsf{X}}_\alpha(\mathrm{d}d|x)\mathbb{P}^{\mathsf{X}|\mathsf{H}}_C(\mathrm{d}x|h)\mu(\mathrm{d}h) && (3.32)\\
&= -\int_H R(\mathbb{P}^{\mathsf{D}|\mathsf{X}}_\alpha, h)\mu(\mathrm{d}h) && (3.33)\\
&= -R_\mu(\mathbb{P}^{\mathsf{D}|\mathsf{X}}_\alpha) && (3.34)
\end{aligned}
$$

□

Complete class theorem

The *complete class theorem* establishes that, under certain conditions, any *admissible* decision rule (Definition 3.2.22) for a see-do model $\mathbb{P}_{C\times H}$ with a utility u must minimise the Bayes risk for a Bayesian model constructed from $\mathbb{P}_{C\times H}$ and a prior over hypotheses $\mu \in \Delta(H)$. This can be interpreted in a similar way to the decision theoretic representation discussed above: if you accept that the relevant assumptions apply to the decision problem at hand, than there is a Bayesian see-do model along with u that captures the important features of this problem. The assumptions are that a see-do model $\mathbb{P}_{C\times H}$ with a utility u that satisfies the relevant conditions is available, and that the principle used to evaluate decision rules should yield an admissible decision rule (though it may also be desired to satisfy other properties as well).

If there are auxiliary requirements for choosing the decision rule, the complete class theorem does not prove that it is easy to find a Bayesian model that will yield rules satisfying these requirements.

See-do models (and statistical decision problems) have a lot of structure – the loss, the assumption that consequences are conditionally independent of observations – that is not actually critical to the complete class theorem. The complete class theorem is a theorem about risk function $R : C \times H \to \mathbb{R}$ that have certain properties. Theorem 3.2.16 shows one way that a risk function can be derived from a see-do model along with a utility. However, it is also possible to derive risk functions from other classes of probability set models with utilities, and if the resulting risk function satisfies the appropriate conditions then the complete class theorem also applies to that class of model. For example, the complete class theorem also applies to see-do models without the assumption that consequences are conditionally independent of observations given the hypothesis and the decision, even though in this case the risk calculation is not the standard calculation for a statistical decision problem.

Definition 3.2.21 (Risk function)**.** Given a set of choice C and a set of hypotheses H, a risk function is a map $R : H \times C \to \mathbb{R}$.

If the second set H were, instead of hypotheses about nature, a set of options available to a second player playing a game, then a "risk function" defines a two-player zero-sum game Ferguson (1967).

Definition 3.2.22 (Admissible choice)**.** Given a risk function $R : C \times H \to \mathbb{R}$, a choice r $\alpha \in C$ dominates a choice $\alpha' \in C$ if for all $h \in H$, $R(\alpha, h) \leq R(\alpha', h)$ and for at least on h^*, $R(\alpha, h) < R(\alpha, h^*)$. An *admissible choice* is a choice $\alpha \in C$ such that there is no $\alpha' \in C$ dominating α.

Definition 3.2.23 (Complete class)**.** A *complete class* is any $B \subset C$ such that, for any $\alpha' \not{i}nB$ there is some $\alpha \in B$ that dominates α'. A *minimal complete class* is a complete class B such that no proper subset of B is complete

Theorem 3.2.24. *If a minimal complete class $B \subset C$ exists then B is the set consisting of all the admissible decision rules.*

Proof. See Ferguson (1967, Theorem 2.1) □

Definition 3.2.25 (Risk set)**.** Given a finite set of hypotheses H, a set of choices C and a risk function $R : C \times H \to \mathbb{R}$, the risk set is the subset of $\mathbb{R}^{|H|}$ given by

$$S := \{(R(\alpha, h))_{h \in H} | \alpha \in C\} \tag{3.35}$$

Theorem 3.2.26 (Complete class theorem)**.** *Given a risk function $R : C \times H \to \mathbb{R}$, if the risk set S is convex, bounded from below and closed downwards, and H is finite, then the set of Bayes optimal choices is a minimal complete class.*

Proof. See Ferguson (1967, Theorem 2.10.2) □

Two examples of the application of the complete class theorem will be presented (Examples 3.2.30 and 3.2.31). In order to explain them, we need a few lemmas.

Lemma 3.2.27. *Given H and C both finite and a risk function $R : C \times H \to \mathbb{R}$ and an associated probability set $\mathbb{P}_C$ on $(\Omega, \mathcal{F})$, Ω finite, if the function*

$$\mathbb{P}_{\alpha,h}^{\mathsf{D}|\mathsf{X}} \mapsto R(\alpha, h) \tag{3.36}$$

is linear and

$$Q := ((\mathbb{P}_{\alpha,h}^{\mathsf{D}|\mathsf{X}})_{h \in H})_{\alpha \in C} \tag{3.37}$$

is convex closed, then the risk set S is convex closed.

Proof. By linearity of

$$\mathbb{P}_{\alpha,h}^{\mathsf{D}|\mathsf{X}} \mapsto R(\alpha, h) \tag{3.38}$$

we also have linearity of

$$(\mathbb{P}_{\alpha,h}^{\mathsf{D}|\mathsf{X}})_{h\in H} \mapsto (R(\alpha, h))_{h\in H} \tag{3.39}$$

Furthermore, Q is bounded when viewed as an element of $\mathbb{R}^{\Omega\times H\times C}$, and so S is the linear image of a compact convex set, and is therefore also compact convex. □

Lemma 3.2.28. *For a see-do model* $(\mathbb{P}_{C\times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D}, \mathsf{H})$ *with utility* $u : X \times Y \to \mathbb{R}$*, the map*

$$\mathbb{P}_{\alpha,h}^{\mathsf{D}|\mathsf{X}} \mapsto R(\alpha, h) \tag{3.40}$$

is linear.

Proof. By definition,

$$\begin{aligned} R(\alpha, h) &= -\mathbb{P}_{\alpha,h}^{\mathsf{XY}} u &(3.41)\\ &= -\mathbb{P}_{C\times\{h\}}^{\mathsf{X}} \odot \mathbb{P}_{\alpha\times h}^{\mathsf{D}|\mathsf{X}} \odot \mathbb{P}_{C\times\{h\}}^{\mathsf{Y}|\mathsf{DX}} u &(3.42)\end{aligned}$$

Which is a composition of kernel products involving $\mathbb{P}_{\alpha\times H}^{\mathsf{D}|\mathsf{X}}$, and kernel products are linear, hence this function is linear. □

The preceding theorem does *not* hold for a utility defined on Ω rather than on $X \times Y$. In this case we have instead

$$-\mathbb{P}_{C\times\{h\}}^{\mathsf{X}} \odot \mathbb{P}_{\alpha\times h}^{\mathsf{D}|\mathsf{X}} \odot \mathbb{P}_{\alpha,h}^{\Omega|\mathsf{DX}} u \tag{3.43}$$

where α appears twice on the right hand side, rendering the map nonlinear.

Lemma 3.2.29. *For finite* X *and* D*, the set of all Markov kernels* $X \rightarrow D$ *is convex closed.*

Proof. From Blackwell (1979), the set of all Markov kernels $X \rightarrow D$ is the convex hull of the set of all deterministic Markov kernels $X \rightarrow D$. There are a finite number of deterministic Markov kernels, and so the convex hull of this set is closed. □

Example 3.2.30. Suppose we have a conditionally independent see-do model $(\mathbb{P}_C, \mathsf{X}, \mathsf{Y}, \mathsf{D}, \mathsf{H})$ along with a bounded utility $u : Y \to \mathbb{R}$ where H, D, X and Y are all finite, and $\{\mathbb{P}_{\alpha}^{\mathsf{D}|\mathsf{X}} | \alpha \in C\}$ is the set of all Markov kernels $X \rightarrow D$. Then the risk set is convex and closed downwards, and so the set of Bayes optimal choices is exactly the set of admissible choices.

The boundedness of the risk set S follows from the boundedness of the utility u; if u is bounded above by k, then S is bounded below in every dimension by $-k$.

The fact that S is convex and closed follows from Lemmas 3.2.27, 3.2.28 and 3.2.29.

Example 3.2.31. As before, but suppose we have the see-do model is not conditionally independent. Because none of the lemmas 3.2.27, 3.2.28 and 3.2.29 made use of the conditional independence assumption, the risk set is still convex and closed downwards and so the set of Bayes optimal choices is also exactly the set of admissible choices.

3.3 Variables

In probability theory, it is standard to assume the existence of a probability space $(\mu, \Omega, \mathcal{F})$ and to define *random variables* as measurable functions from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. However, variables aren't *just* functions – they're also typically understood to correspond to some measured aspect of the real world. For example, Pearl (2009) offers the following two, purportedly equivalent, definitions of variables:

> By a *variable* we will mean an attribute, measurement or inquiry that may take on one of several possible outcomes, or values, from a specified domain. If we have beliefs (i.e., probabilities) attached to the possible values that a variable may attain, we will call that variable a random variable.
>
> This is a minor generalization of the textbook definition, according to which a random variable is a mapping from the sample space (e.g., the set of elementary events) to the real line. In our definition, the mapping is from the sample space to any set of objects called "values," which may or may not be ordered.

However, these are actually two different things. The first is a *measurement*, which is something we can do in the real world that produces as a result an element of a mathematical set. The second is a *function*, a purely mathematical object with a domain and a codomain and a mapping from the former into the latter. Measurement procedures play the extremely important role of "pointing to the parts of the world" that the model addresses.

The general scheme considered in this work is to assume that there is a collection of "complete measurement procedure" $\mathcal{S}_\alpha$, one for each choice $\alpha \in C$. $\mathcal{S}_\alpha$ is considered to be the procedure that measures all quantities of interest, and any subprocedure corresponding to a particular quantity of interest reconstructed from the result of $\mathcal{S}$ by applying a function to its result. The function X that, when applied to the result of $\mathcal{S}$, yields the result of a measurement subprocedure $\mathcal{X}$ is the *variable* associated with the measurement procedure $\mathcal{X}$. In this way, a variable X – which is by itself just a mathematical function – is associated with a measurement procedure in the real world.

3.3.1 Variables and measurement procedures

Consider Newton's second law in the form $\mathsf{F} = \mathsf{MA}$. This model relates "variables" F, M and A. As Feynman (1979) noted, in order to understand this law, some pre-existing understanding of force, mass and acceleration is required. In order to offer a numerical value for the net force on a given object is, even the most knowledgeable physicist will have to go and do a measurement, which involves interacting with the object in some manner that cannot be completely mathematically specified, and which will return a numerical value that will be taken to be the net force.

In order to make sense of the equation $\mathsf{F} = \mathsf{MA}$, it must be understood relative to some measurement procedure $\mathcal{S}$ that simultaneously measures the force on an object, its mass and its acceleration, which can be recovered by the functions F, M and A respectively. The equation then says that, whatever result s this procedure yields, $\mathsf{F}(s) = \mathsf{M}(s)\mathsf{A}(s)$ will hold.

A measurement procedure $\mathcal{S}$ is akin to Menger (2003)'s notion of variables as "consistent classes of quantities" that consist of pairing between real-world objects and quantities of some type. $\mathcal{S}$ itself is not a well-defined mathematical thing. At the same time, the set of values it may yield *is* a well-defined mathematical set. No actual procedure can be guaranteed to return elements of a mathematical set known in advance – anything can fail – but we assume that we can study procedures reliable enough that we don't lose much by ignoring this possibility.

Note that, because $\mathcal{S}$ is not a purely mathematical thing, we cannot perform mathematical reasoning with $\mathcal{S}$ directly. It is much more practical to relegate $\mathcal{S}$ to the background, and reason in terms of the functions F, M and A. However, even if we don't talk about it much, $\mathcal{S}$ remains an important element of the law.

3.3.2 Measurement procedures

Definition 3.3.1 (Measurement procedure)**.** A *measurement procedure* $\mathcal{B}$ is a procedure that involves interacting with the real world somehow and delivering an element of a mathematical set X as a result. A procedure $\mathcal{B}$ is said to takes values in a set B.

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

Definition 3.3.2 (Values yielded by procedures)**.** $\mathcal{B} \bowtie x$ is the proposition that the the procedure $\mathcal{B}$ will yield the value $x \in X$. $\mathcal{B} \bowtie A$ for $A \subset X$ is the proposition $\vee_{x\in A}\mathcal{B} \bowtie x$.

Definition 3.3.3 (Equivalence of procedures)**.** Two procedures $\mathcal{B}$ and $\mathcal{C}$ are equal if they both take values in X and $\mathcal{B} \bowtie x \iff \mathcal{C} \bowtie x$ for all $x \in X$.

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

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

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

Definition 3.3.4 (Composition of functions with procedures)**.** Given a procedure $\mathcal{B}$ that takes values in some set B, and a function $f : B \to C$, define the "composition" $f \circ \mathcal{B}$ to be any procedure $\mathcal{C}$ that yields $f(x)$ whenever $\mathcal{B}$ yields x. We can construct such a procedure by describing the steps: first, do $\mathcal{B}$ and secondly, apply f to the value yielded by $\mathcal{B}$.

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

$$
\begin{aligned}
(g \circ f) \circ \mathcal{B} \text{ yields } x &\iff B \text{ yields } (g \circ f)^{-1}(x) && (3.44)\\
&\iff B \text{ yields } f^{-1}(g^{-1}(x)) && (3.45)\\
&\iff f \circ B \text{ yields } g^{-1}(x) && (3.46)\\
&\iff g \circ (f \circ B) \text{ yields } x && (3.47)
\end{aligned}
$$

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

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

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

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

3.3.3 Observable variables

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

Definition 3.3.5 (Observable variable)**.** Given a measurement procedure $\mathcal{S}$ taking values in $(\Psi, \mathcal{E})$, an observable variable is a pair $(\mathsf{X} \circ \mathcal{S}, \mathsf{X})$ where $\mathsf{X} : (\Psi, \mathcal{E}) \to (X, \mathcal{X})$ is a measurable function and $\mathcal{X} := \mathsf{X} \circ \mathcal{S}$ is the measurement procedure induced by X and $\mathcal{S}$.

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

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

Definition 3.3.6 (Event)**.** Given the complete procedure $\mathcal{S}$ taking values in Ψ and an observable variable $(\mathsf{X} \circ \mathcal{S}, \mathsf{X})$ for $\mathsf{X} : \Psi \to X$, the *event* $\mathsf{X} \bowtie x$ is the set $\mathsf{X}^{-1}(x)$ for any $x \in X$.

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

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

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

3.3.4 Model variables

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

A *model variable* is simply a measurable function with domain $(\Psi, \mathcal{E})$.

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

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

Definition 3.3.7 (Unobserved variable)**.** Given a sample space $(\Omega, \mathcal{F})$ and a complete observable variable $\mathsf{S} : \Omega \to \Psi$, a model variable $\mathsf{Y} : \Omega \to Y$ is *unobserved* if $\mathsf{Y}(\mathsf{S} \bowtie s) = Y$ for all $s \in \Psi$.

3.3.5 Variable sequences and partial order

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

Define the partial order on variables $\preccurlyeq$ where $\mathsf{X} \preccurlyeq \mathsf{Y}$ can be read "X is completely determined by Y".

Definition 3.3.8 (Variables determined by another variable)**.** Given a sample space $(\Omega, \mathcal{F})$ and variables $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$, $\mathsf{X} \preccurlyeq \mathsf{Y}$ if there is some $f : Y \to X$ such that $\mathsf{X} = f \circ \mathsf{Y}$.

Clearly, $\mathsf{X} \preccurlyeq (\mathsf{X}, \mathsf{Y})$ for any X and Y.

3.3.6 Decision procedures

The kind of problem we want to solve requires us to compare the consequences of different choices from a set of possibilities C. We take the *consequences of* $\alpha \in C$ to refer to the values obtained by some measurement procedure $\mathcal{S}_\alpha$ associated with the choice α.

As we have said, what exactly a "measurement procedure" is is a bit vague – it's "what we actually do to get the numbers we associate with variables". It seems we could describe the above in terms of a single measurement procedure $\mathcal{S}$, which involves:

1. Choose α

2. Proceed according to $\mathcal{S}_\alpha$

However, $\mathcal{S}$ is problematic to model. The model is often part of the process of choosing α, and so a model of $\mathcal{S}$ that involves the step "choose α" will be self-referential. Because of this, we don't try to model $\mathcal{S}$, and whether this changes anything is an open question.

Definition 3.3.9 (Decision procedure)**.** A decision procedure is a collection $\{\mathcal{S}_\alpha\}_{\alpha\in C}$ of measurement procedures.

Like measurement procedures, a decision procedure $\{\mathcal{S}_\alpha\}_{\alpha\in A}$ isn't a well-defined mathematical object; it's not really a "set", because the contents are real-world actions.

Chapter 4

Models of repeatable decision problems

Chapter 2 introduced probability sets as generic tools for causal modelling, while Chapter 3 examined how probability set models can be used in decision problems and Section 3.2.5 in particular introduced *see-do* models, which featured variables representing observations and consequences, choices and hypotheses. So far, no attention has been paid to how exactly anybody might use a see-do model to help them "learn from observations". Observations and consequences in see-do models could be just about anything – they need not take values in the same set, nor be related to one another in any particular way. While there might be interesting problems involving models where observations and consequences are different kinds of things, causal inference in practice is usually concerned with problems where a sequence of observations is available and the consequences of interest are of the "same type" as observations. Such problems involve *repeatable* measurement procedures - the procedure can be broken down into a sequence sub-procedures that all have something in common with one another. This chapter aims to better understand the "something in common" that these sub-procedures share.

Repeatable procedures in classical statistics are associated with the assumption of *exchangeability of observations*. Exchangeable models express the assumption that the results of sub-procedures can be swapped without changing the problem in any important way – this is the "something in common" that exchangeable sub-procedures have (note that in a non-Bayesian setting, this assumption is called permutability and is weaker than exchangeability, see Walley (1991, pg. 463)). Decision problem models usually aren't compatible with this assumption; often, the decision maker's choices will affect some sub-procedures in different ways to others. For example, the decision maker might choose to record 10 values without interaction, then tweak the system of interest before recording their 11th value – in this case, swapping the results of the 11th sub-procedure with the 1st *does* change the problem in an important way, because it changes the rank order of the "tweaked" sub-procedure. The need to account for different possible "inputs" to each sub-procedure makes the question of commonality between these sub-procedures substantially more complicated than the assumption of exchangeability.

This chapter introduces *conditionally independent and identical response functions* as the "something in common" that a sequence of subprocedures might have. Section 4.2 defines conditionally independent and identical response functions and explores different examples of models that feature them. The assumption of conditionally independent and identical response functions can be combined with additional assumptions, and these yield different classes of models.

Section 4.3 shows that a Markov kernel can be represented as a mixture of conditionally independent and identical response functions if and only if it is *causally contractible* (Theorem 4.3.11). Additionally, Theorem 4.3.13 shows that a Markov kernel that satisfies the weaker condition under an additional condition can be represented as a mixture of conditionally independent and identical response functions where the mixture depends on a symmetric function of the entire sequence of inputs.

Probability set models with conditionally independent and identical response functions all feature causally contractible Markov kernels, but depending on the class of model under consideration, these Markov kernels can play different roles. The simplest case is where each input is assumed to be independent of all data that has come previously and the hypothesis (Definition 4.2.1) is assumed independent of the inputs. A model satisfying these assumptions features a causally contractible uniform conditional probability. Such models are explored in Section 4.3.3. Relaxing the assumption of independence between the hypothesis and the inputs yields a causally contractible *uniform 2-comb* (Definition 4.5.1). This is explored in Section 4.4. Finally, we explore models where inputs can depend on previous data in Section 4.5, in which case the model features a causally contractible *uniform n-comb*.

4.1 Previous work

This chapter draws on three different lines of work. The first is the study of representations of symmetric of probability models. The equivalence between infinite exchangeable probability models and mixtures of independent and identically distributed models was shown by de Finetti ([1937] 1992). This result has been extended in many ways, including to finite sequences Kerns and Székely (2006); Diaconis and Freedman (1980) and for partially exchangeable arrays Aldous (1981). A comprehensive overview of results is presented in Kallenberg (2005b). Particularly similar to our result is the notion of "partial exchangeability" from Diaconis (1988).

The second line of work is the study of exchangeability-like assumptions in causal models. Lindley and Novick (1981) discuss models consisting of a sequence of exchangeable observations along with "one more observation", a structure that is similar to the models with observations and consequences discussed in section 5.1. Lindley discusses the application of this model to questions of causation, but does not explore this deeply due to the perceived difficulty of finding a satisfactory definition of causation. Rubin (2005)'s overview of causal inference with potential outcomes along with the text Imbens and Rubin (2015) make use of the assumption of exchangeable potential outcomes to prove several identification results. Saarela et al. (2020), using structural causal models, proposes *conditional exchangeability*, which refers to the invariance of a joint distribution over outcomes under "surgical switches" of the values of causal variables of interest. This definition depends on having a structural model, a property not shared by the current work.

Exchangeability in the setting of causal models is often discussed in terms of the exchangeability of *people* (or more generically, *experimental units*). Hernán (2012); Greenland and Robins (1986); Banerjee et al. (2017); Dawid (2020) all discuss assumptions along these lines.

A stronger assumption than commutativity of exchange is *causal contractibility* (Definition 4.3.1), which adds the assumption of *locality*. This additional assumption appears to have similarities to the stable unit treatment distribution assumption (SUTDA) in Dawid (2020), and the stable unit treatment value assumption (SUTVA) in (Rubin, 2005): "("SUTVA) comprises two subassumptions. First, it assumes that *there is no interference between units (Cox 1958)*; that is, neither $Y_i(1)$ nor $Y_i(0)$ is affected by what action any other unit received. Second, it assumes that *there are no hidden versions of treatments*; no matter how unit i received treatment 1, the outcome that would be observed would be $Y_i(1)$ and similarly for treatment 0.

Finally, the idea of *combs* in probabilistic models was first proposed by Chiribella et al. (2008) and an application to causal models was developed by Jacobs et al. (2019).

4.2 Conditionally independent and identical response functions

Suppose a decision maker is implementing a decision procedure where they'll make a choice and subsequently receive a sequence of paired values $(\mathcal{D}_i, \mathcal{Y}_i)$, with their objective depending on the output values yielded by $\mathcal{Y}_i$s only. Usually the $\mathcal{D}_i$s, which we call "inputs", are under the decision maker's control to some extent, but this might not always be the case. For example, perhaps the first m pairs come from data collected by someone else, where the decision maker has no control over inputs, and the next n depend on their own actions where they have complete control over the inputs.

Suppose the decision maker uses a probability set $\mathbb{P}_C$ to model such a procedure, and variables $(\mathsf{D}_i, \mathsf{Y}_i)$ are associated with the inputs and outputs. There are two different relationships between D_i and Y_i that might be of interest to the decision maker:

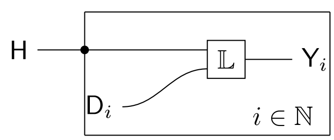
- For some choice α, $j > m$ and some fixed value of D_j, what are the *likely consequences* with regard to Y_j?
- For some choice α, all $i \leq m$ with some fixed value of D_i, what is the *relative frequency* of different values of Y_i?

The first is what the decision maker wants to know in order to make a good decision, and the second is something they can learn from the data before taking any actions. In particular, if the decision maker has a good reason to thing that the two relationships should be (approximately) the same, then they may choose to set (or influence) D_j for $j > m$ to whatever value appears the most favourable according to the preceding data.

More precisely, we are interested in models $\mathbb{P}_C$ where the probabilistic relationship between each D_i and the corresponding Y_i is unknown but identical for all indices i. To model this, we introduce a hypothesis H that represents this unknown relationship, and assert that the distribution of Y_i given $(\mathsf{D}_i, \mathsf{H})$ is identical for all i. Refer to Section 2.3.2 for the definition of the plate notation.

Definition 4.2.1 (Conditionally independent and identical response functions)**.** A probability set $\mathbb{P}_C$ on $(\Omega, \mathcal{F})$ with variables $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$ and $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ has *conditionally independent and identical response functions* if there is some hypothesis $\mathsf{H} : \Omega \to H$ and $\mathbb{L} : H \times D \rightarrow Y$ such that

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{DH}} \overset{\mathbb{P}_C}{\cong} \qquad (4.1)$$

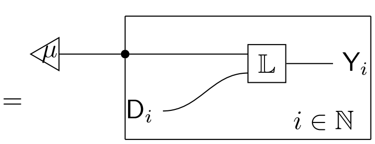


$$\iff \qquad (4.2)$$

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{DH}}(\bigtimes_{i\in\mathbb{N}} A|d,h) = \prod_{i=1}^{\infty} \mathbb{L}(A_i|d_i,h) \qquad \forall A_i \in \mathcal{Y}, h \in H, d \in D^{\mathbb{N}} \qquad (4.3)$$

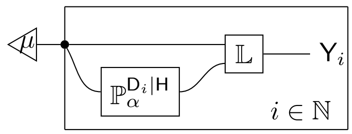
Definition 4.2.1 implies $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{D}_{\neq i}, \mathsf{Y}_{\neq i}, \mathsf{C})|(\mathsf{D}_i, \mathsf{H})$. We can consider a number of different kinds of models where this holds. A simple family of models can be obtained by "attaching" a prior μ over H to Equation 4.1. This yields the Markov kernel $\mathbb{K} : D^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$

$$\mathbb{K} := \qquad (4.4)$$



$\mathbb{K} = \mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ if $\mathsf{H} \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{D}, \mathsf{C})$ and $\mathsf{D}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}_{[i-1]}|(\mathsf{D}_{[i-1]}, \mathsf{C})$. We say that such models have an *input-independent hypothesis* and *data-independent inputs*, and they are considered in more detail in Section 4.3.3. This is a limited family of models, but it can be a helpful starting point to understand more general ones. Theorem 4.3.15 establishes that given a data-independent model $\mathbb{P}_C$, $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ is *causally contractible* if and only if it can be represented in the form given by Equation 4.4.

Dropping the assumption $\mathsf{H} \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{D}, \mathsf{C})$ yields models that simply have data-independent inputs. In this case, we can write for any $\alpha \in C$

$\mathbb{P}_\alpha^{\mathsf{YD}} =$ 

(4.5)

Of particular interest is the case where $\mathbb{P}_\alpha^{\mathsf{D}_i|\mathsf{H}} \neq \mathbb{P}_\alpha^{\mathsf{D}_j|\mathsf{H}}$ for some $i \neq j$. In this case, we have a model with *varying inputs*, and this class of models is explored in Section 4.4. For models with varying inputs, the 2-comb $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}\leftarrow\mathsf{H}}$ is causally contractible (see Section 4.5.1 for an explanation of the notation), and this suggests a symmetry with respect to "input switching" (which we explain in the section). Models with varying inputs are particularly closely related to many standard causal inference problems, as we will see in Chapter 5.

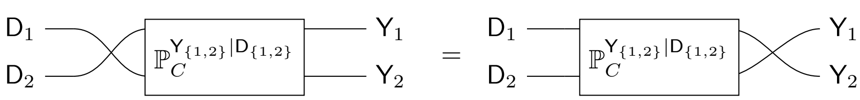
Dropping the assumption $\mathsf{D}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}_{\neq i}|(\mathsf{C}, \mathsf{H})$ yields models with *data-dependent inputs*. An example of a model in this class is the multi-armed bandit, a standard toy problem in the field of reinforcement learning Barto (1998). Models with data-dependent inputs have causally contractible $\mathbb{N}$-combs $\mathbb{P}_C^{\mathsf{Y}\wr\mathsf{D}}$ (see Section 4.5.1 again for an explanation of the notation).

4.3 Causally contractible Markov kernels - definitions and explanation

In this section we prove representation theorems for Markov kernels that satisfy the assumptions of causal contractibility and exchange commutativity. These theorems will be applied to various families of probability set models in the following sections (specifically, Section 4.3.3, Section 4.4 and Section 4.5).

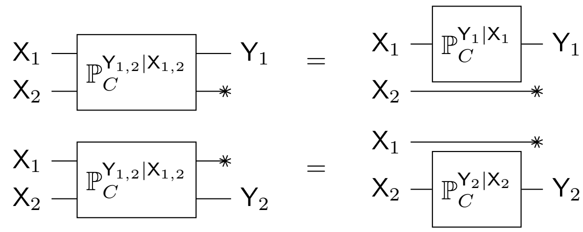
The assumptions of exchange commutativity and *locality* together make causal contractibility. Exchange commutativity expresses a sense in which a Markov kernel $\mathbb{K} : X^{\mathbb{N}} \to Y^{\mathbb{N}}$ can be symmetric: if it yields the same result from shuffling inputs as it does from shuffling outputs. Locality is the assumption that subsets of outputs are probabilistically independent of the non-corresponding inputs, conditioned on the corresponding inputs.

Graphical notation can offer an intuitive picture of these two assumptions. In the simplified case of a sequence of length 2 (that is, $\mathbb{K} : X^2 \to Y^2$), exchange commutativity for two inputs and outputs is given by the following equality:



(4.6)

swapping the inputs is equivalent to applying the same swap to the outputs. Locality is given by the following pair of equalities:



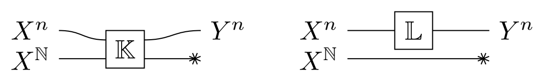
(4.7)

(4.8)

and expresses the idea that the outputs are independent of the non-corresponding input, conditional on the corresponding input.

The general definitions follow.

Definition 4.3.1 (Locality)**.** A Markov kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is *local* if for all $n \in \mathbb{N}$, $A_i \in \mathcal{Y}$, $(x_{[n]}, x_{[n]^C}) \in \mathbb{N}$ there exists $\mathbb{L} : X^n \rightarrow Y^n$ such that



$$= \tag{4.9}$$

$$\iff \tag{4.10}$$

$$\mathbb{K}(\bigtimes_{i\in[n]} A_i \times Y^{\mathbb{N}}|x_{[n]}, x_{[n]^C}) = \mathbb{L}(\bigtimes_{i\in[n]} A_i|x_{[n]}) \tag{4.11}$$

Definition 4.3.2 (Exchange commutativity)**.** A Markov kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ *commutes with exchange* if for all finite permutations $\rho : \mathbb{N} \to \mathbb{N}$, $A_i \in \mathcal{Y}$, $(x_{[n]}, x_{[n]^C}) \in \mathbb{N}$

$$\mathbb{K}\text{swap}_{\rho,Y} = \text{swap}_{\rho,X}\mathbb{K} \tag{4.12}$$

$$\iff \tag{4.13}$$

$$\mathbb{K}(\bigtimes_{i\in\mathbb{N}} A_{\rho(i)}|(x_i)_{i\in\mathbb{N}}) = \mathbb{K}(\bigtimes_{i\in\mathbb{N}} A_i|(x_{\rho(i)})_{i\in\mathbb{N}}) \tag{4.14}$$

Causal contractibility is the conjunction of both assumptions.

Definition 4.3.3 (Causal contractibility)**.** A Markov kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is *causally contractible* if it is local and commutes with exchange.

4.3.1 Properties of causally contractible Markov kernels

A causally contractible Markov kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ can be mapped to "contracted" versions of the kernel using the del map. Theorem 4.3.7 establishes that any two contraction of $\mathbb{K}$ with equivalent codomains yields the same result (see also Theorem 4.3.16). This feature is the motivation for the name *causal contractibility.*

Theorem 4.3.8 shows that exchange commutativity and locality are independent assumptions.

Before these theorems are proved, the following definition and Lemma will prove helpful.

All swaps can be written as a product of transpositions, so proving that a property holds for all finite transpositions is enough to show it holds for all finite swaps. It's useful to define a notation for transpositions.

Definition 4.3.4 (Finite transposition)**.** Given two equally sized sequences $A = (a_i)_{i\in[n]}$, $B = (b_i)_{i\in[n]}$, $A \leftrightarrow B : \mathbb{N} \to \mathbb{N}$ is the permutation that sends the ith element of A to the ith element of B and vise versa. Note that $A \leftrightarrow B$ is its own inverse.

Lemma 4.3.5 is used to extend finite sequences to infinite ones, and is used in a number of upcoming theorems.

Lemma 4.3.5 (Infinitely extended kernels)**.** *Given a collection of Markov kernels $\mathbb{K}_i : X^i \rightarrowtriangle Y^i$ for all $i \in \mathbb{N}$, if we have for every $j > i$*

$$\mathbb{K}_j(id_{X_i} \otimes del_{X_{j-i}}) = \mathbb{K}_i \otimes del_{X_{j-i}} \tag{4.15}$$

then there is a unique Markov kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrowtriangle Y^{\mathbb{N}}$ such that for all $i, j \in \mathbb{N}$,$j > i$

$$\mathbb{K}(id_{X_i} \otimes del_{X_{j-i}}) = \mathbb{K}_i \otimes del_{X_{j-i}} \tag{4.16}$$

Proof. Take any $x \in X^{\mathbb{N}}$ and let $x_{|m} \in X^n$ be the first n elements of x. By Equation 4.15, for any $A_i \in \mathcal{Y}$, $i \in [m]$

$$\mathbb{K}_n(\bigtimes_{i\in[m]} A_i \times Y^{n-m}|x_{|n}) = \mathbb{K}_m(\bigtimes_{i\in[m]} A_i|x_{|m}) \tag{4.17}$$

Furthermore, by the definition of the swap map for any permutation $\rho : [n] \to [n]$

$$\mathbb{K}_n\text{swap}_\rho(\bigtimes_{i\in[m]} A_{\rho(i)} \times Y^{n-m}|x_{|n}) = \mathbb{K}_n(\bigtimes_{i\in[m]} A_i \times Y^{n-m}|x_{|n}) \tag{4.18}$$

thus by the Kolmogorov Extension Theorem (Çinlar, 2011), for each $x \in X^{\mathbb{N}}$ there is a unique probability measure $\mathbb{Q}_x \in \Delta(Y^{\mathbb{N}}$ satisfying

$$\mathbb{Q}_d(\bigtimes_{i\in[n]} A_i \times Y^{\mathbb{N}}) = \mathbb{K}_n(\bigtimes_{i\in[n]} A_{\rho(i)}|d_{|n}) \tag{4.19}$$

Furthermore, for each $\{A_i \in \mathcal{Y}|i \in \mathbb{N}\}$, $n \in \mathbb{N}$ note that for $p > n$

$$\mathbb{Q}_d(\bigtimes_{i\in[n]} A_i \times Y^{\mathbb{N}}) \geq \mathbb{Q}_d(\bigtimes_{i\in[p]} A_i \times Y^{\mathbb{N}}) \tag{4.20}$$

$$\geq \mathbb{Q}_d(\bigtimes_{i\in\mathbb{N}} A_i) \tag{4.21}$$

so by the Monotone convergence theorem, the sequence $\mathbb{Q}_d(\bigtimes_{i\in[n]} A_i)$ converges as $n \to \infty$ to $\mathbb{Q}_d(\bigtimes_{i\in\mathbb{N}} A_i)$. $d \mapsto \mathbb{Q}_d^{\mathsf{Z}_n}(\bigtimes_{i\in[n]} A_i)$ is measurable for all n, $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$ by Equation 4.19, and so $d \mapsto Q_d$ is also measurable.

Thus $d \mapsto Q_d$ is the desired $\mathbb{P}_C^{\mathsf{Y}_\mathbb{N} | \mathsf{D}_\mathbb{N}} : D^\mathbb{N} \rightarrow Y^\mathbb{N}$. □

Theorem 4.3.7 shows that, given a causally contractible kernel, the following operations yield equivalent results:

- Marginalising all but the first n outputs
- Marginalising all outputs except for the positions $A \subset \mathbb{N}$ where $|A| = n$, and swapping the first n inputs with the elements of A

Definition 4.3.6 (Marginalising kernel)**.** Given $(X, \mathcal{X})$ and $A \subset \mathbb{N}$, $\mathrm{marg}_A : X^\mathbb{N} \rightarrow X^A$ is the Markov kernel given by

$$\bigotimes_{i\in\mathbb{N}} \mathrm{switch}_{A,i} \tag{4.22}$$

where

$$\mathrm{switch}_A = \begin{cases} \mathrm{id}_X & i \in A \\ \mathrm{del}_X & i \notin A \end{cases} \tag{4.23}$$

Theorem 4.3.7 (Equality of equally sized contractions)**.** *A Markov kernel $\mathbb{K} : X^\mathbb{N} \rightarrow Y^\mathbb{N}$ is* causally contractible *if and only if for every $n \in \mathbb{N}$ and every $A \subset \mathbb{N}$ there exists some $\mathbb{L} : X^n \rightarrow Y^n$ such that*

$$\mathbb{K} marg_A = swap_{[n]\leftrightarrow A} \mathbb{L} \otimes del_{X^\mathbb{N}} \tag{4.24}$$

Proof. Only if: By exchange commutativity

$$\mathrm{swap}_{[n]\leftrightarrow A} \mathbb{K} = \mathbb{K} \mathrm{swap}_{[n]\leftrightarrow A} \tag{4.25}$$

multiply both sides by $\mathrm{swap}_{[n]\leftrightarrow A}$ on the right and, because $\mathrm{swap}_{[n]\leftrightarrow A}$ is its own inverse,

$$\mathrm{swap}_{[n]\leftrightarrow A} \mathbb{K} \mathrm{swap}_{[n]\leftrightarrow A} = \mathbb{K} \tag{4.26}$$

so

$$\begin{aligned} \mathbb{K}\mathrm{marg}_A &= \mathrm{swap}_{[n]\leftrightarrow A} \mathbb{K} \mathrm{swap}_{[n]\leftrightarrow A} \mathrm{marg}_A && (4.27) \\ &= \mathrm{swap}_{[n]\leftrightarrow A} \mathbb{K} \mathrm{marg}_{[n]} && (4.28) \end{aligned}$$

By locality, there exists some $\mathbb{L} : X^n \rightarrow Y^n$ such that

$$\begin{aligned} \mathbb{K}\mathrm{marg}_{[n]} &= \mathbb{K}(\mathrm{id}_{[n]} \otimes \mathrm{del}_{X^\mathbb{N}}) && (4.29) \\ &= \mathbb{L} \otimes \mathrm{del}_{X^\mathbb{N}} && (4.30) \end{aligned}$$

If: Taking $A = [n]$ for all n establishes locality.

For exchange commutativity, note that for all $x \in X^{\mathbb{N}}$, $n \in \mathbb{N}$, we have

$$\begin{aligned}
\mathrm{swap}_{A\leftrightarrow[n]}\mathbb{K}\mathrm{marg}_A &= \mathrm{swap}_{A\leftrightarrow[n]}\mathbb{K}\mathrm{swap}_{A\leftrightarrow[n]}(\mathrm{id}_{[n]} \otimes \mathrm{del}_{X^{\mathbb{N}}}) && (4.31)\\
&= \mathbb{K}\mathrm{marg}_{[n]} && (4.32)\\
&= \mathbb{K}(\mathrm{id}_{[n]} \otimes \mathrm{del}_{X^{\mathbb{N}}}) && (4.33)
\end{aligned}$$

Then by Lemma 4.3.5

$$\mathrm{swap}_{A\leftrightarrow[n]}\mathbb{K}\mathrm{swap}_{A\leftrightarrow[n]} = \mathbb{K} \tag{4.34}$$

Consider an arbitrary finite permutation $\rho : \mathbb{N} \to \mathbb{N}$. ρ can be decomposed into a finite set of cyclic permutations on disjoint orbits. Each cyclic permutation is simply the composition of some set of transpositions, and so ρ itself can be written as a composition of a sequence of transpositions. Thus for any finite $\rho : \mathbb{N} \to \mathbb{N}$

$$\mathrm{swap}_\rho\mathbb{K}\mathrm{swap}_\rho = \mathbb{K} \tag{4.35}$$

□

Theorem 4.3.8 shows that neither locality nor exchange commutativity is implied by the other.

Theorem 4.3.8. *Exchange commutativity does not imply locality or vise versa.*

Proof. First, a Markov kernel that exhibits exchange commutativity but not locality. Suppose $D = Y = \{0, 1\}$ and $\mathbb{K} : D^2 \rightarrowtriangle Y^2$ is given by

$$\mathbb{K}(y_1, y_2|d_1, d_2) = [\![(y_1, y_2) = (d_1 + d_2, d_1 + d_2)]\!] \tag{4.36}$$

then

$$\mathbb{K}(y_1, Y|d_1, d_2) = [\![y_1 = d_1 + d_2]\!] \tag{4.37}$$

and there is no function depending on y_1 and d_1 only that is equal to this. Thus $\mathbb{K}$ does not satisfy locality.

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

$$\begin{aligned}
\mathrm{swap}_{\rho,D}\mathbb{K}(y_1, y_2|d_1, d_2) &= \mathbb{K}(y_1, y_2|d_2, d_1) && (4.38)\\
&= [\![(y_1, y_2) = (d_2 + d_1, d_2 + d_1)]\!] && (4.39)\\
&= [\![(y_1, y_2) = (d_1 + d_2, d_1 + d_2)]\!] && (4.40)\\
&= [\![(y_2, y_1) = (d_1 + d_2, d_1 + d_2)]\!] && (4.41)\\
&= \mathbb{K}\mathrm{swap}_{\rho,Y}(y_1, y_2|d_1, d_2) && (4.42)
\end{aligned}$$

so $\mathbb{K}$ commutes with exchange.

Next, a Markov kernel that satisfies locality but does not commute with exchange. Suppose again $D = Y = \{0,1\}$ and $\mathbb{K} : D^2 \rightarrow Y^2$ is given by

$$\mathbb{K}(y_1, y_2|d_1, d_2) = [\![(y_1, y_2) = (0,1)]\!] \tag{4.43}$$

Then:

$$\begin{aligned}
\mathbb{K}(y_1|d_1, d_2) &= [\![y_1 = 0]\!] && (4.44)\\
&= \mathbb{K}(y_1|d_1) && (4.45)\\
\mathbb{K}(y_2|d_1, d_2) &= [\![y_2 = 1]\!] && (4.46)\\
&= \mathbb{K}(y_2|d_2) && (4.47)
\end{aligned}$$

so $\mathbb{K}$ satisfies locality.

However, $\mathbb{K}$ does not commute with exchange.

$$\begin{aligned}
\mathrm{swap}_{\rho(\mathsf{D})}\mathbb{K}(y_1, y_2|d_1, d_2) &= \mathbb{K}(y_1, y_2|d_2, d_1) && (4.48)\\
&= [\![(y_1, y_2) = (0,1)]\!] && (4.49)\\
&\neq [\![(y_2, y_1) = (0,1)]\!] && (4.50)\\
&= \mathbb{K}\mathrm{swap}_{\rho(\mathsf{D})}(y_1, y_2|d_1, d_2) && (4.51)
\end{aligned}$$

□

Theorem 4.3.8 presents abstract counterexamples to show that the assumptions of exchange commutativity and locality are independent. For some more practical examples, a model of the treatment of several patients who are known to have different illnesses might satisfy consequence locality but not exchange commutativity. Patient B's treatment can be assumed not to affect patient A, but the same results would not be expected from giving patient A's treatment to patient B as from giving patient A's treatment to patient A.

A model of strategic behaviour might satisfy exchange commutativity but not locality. Suppose a decision maker is observing people playing a game where they press a red or green button, and (for reasons mysterious to the decision maker), receive a payout randomly of 0 or \$100. The decision maker might reason that the results should be the same no matter who presses a button, but also that people will be more likely to press the red button if the red button tends to give a higher payout. In this case, the decision maker's prediction for the payout of the ith attempt given the red button has been pressed will be higher if the proportion of red button presses in the entire dataset is higher. There are other reasons why exchange commutativity might hold but not locality – Dawid (2000) offers the alternative example of herd immunity in vaccination campaigns. In this case, the overall proportion of the population vaccinated will affect the disease prevalence over and above an individual's vaccination status.

As an aside, although locality could be described as an assumption that there is no interference between inputs and outputs of different indices, it actually allows for some models with certain kinds of interference between actions and outcomes of different indices. For example: consider an experiment where I first flip a coin and record the results of this flip as the outcome of the "step 1". Subsequently, I can choose either to copy the outcome from step 1 to be the input for "step 2" (this is the choice $\mathsf{D}_1 = 0$), or flip a second coin use this as the input for step 2 (this is the choice $\mathsf{D}_1 = 1$). At the second step, I may further choose to copy the provisional results ($\mathsf{D}_2 = 0$) or invert them ($\mathsf{D}_2 = 1$). Then

$$\mathbb{P}_S^{\mathsf{Y}_1|\mathsf{D}}(y_1|d_1, d_2) = 0.5 \tag{4.52}$$

$$\mathbb{P}_S^{\mathsf{Y}_2|\mathsf{D}}(y_2|d_1, d_2) = 0.5 \tag{4.53}$$

- The marginal distribution of both experiments in isolation is Bernoulli(0.5) no matter what choices I make, so a model of this experiment would satisfy Definition 4.3.1
- Nevertheless, the choice at step 1 affects the result of step 2

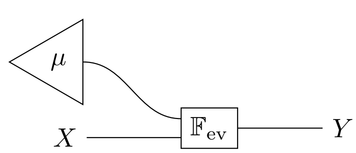
4.3.2 Representation theorems for causally contractible Markov kernels

Theorem 4.3.9 shows that a causally contractible Markov kernel can be represented as the product of a column exchangeable probability distribution and a "lookup function". This representation is identical to the representation of potential outcomes models (see, for example, Rubin (2005)), but Theorem 4.3.9 applies to arbitrary kernels and the resulting representation will usually not be interpretable as a potential outcomes models. This theorem allows De Finetti's theorem to be applied to the column exchangeable probability distribution, which is a key step in proving the main result (Theorem 4.3.11).

Theorem 4.3.13 then extends Theorem 4.3.11 to the case of a Markov kernel with commutativity of exchange only. In this case the latent conditioning variable may depend on a symmetric function of the inputs, assuming the distribution of inputs is dominated by some exchangeable distribution.

Theorem 4.3.9. *A Markov kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is causally contractible if and only if there exists a column exchangeable probability distribution $\mu\Delta(Y^{|X|\times\mathbb{N}})$ such that*

$\mathbb{K} =$



(4.54)

$$\iff \tag{4.55}$$

$$\mathbb{K}(A|(x_i)_{i\in\mathbb{N}}) = \mu\Pi_{(x_i i)_{i\in\mathbb{N}}}(A)\forall A \in \mathcal{Y}^{\mathbb{N}} \tag{4.56}$$

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

$$(y_{ji})_{j,i\in X\times\mathbb{N}} \mapsto (y_{d_i i})_{i\in\mathbb{N}} \tag{4.57}$$

that projects the (x_i, i) *indices of* y *for all* $i \in \mathbb{N}$*, and* $\mathbb{F}_{ev}$ *is the Markov kernel associated with the evaluation map*

$$ev : X^{\mathbb{N}} \times Y^{X\times\mathbb{N}} \to Y \tag{4.58}$$

$$((x_i)_{\mathbb{N}}, (y_{ji})_{j,i\in X\times\mathbb{N}}) \mapsto (y_{x_i i})_{i\in\mathbb{N}} \tag{4.59}$$

Proof. Only if: Choose $e := (e_i)_{i\in\mathbb{N}}$ such that $e_{i+|X|j}$ is the ith element of X for all $i, j \in \mathbb{N}$.

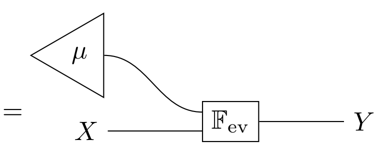
Define

$$\mu(\bigtimes_{(i,j)\in X\times\mathbb{N}} A_{ij}) := \mathbb{K}(\bigtimes_{(i,j)\in X\times\mathbb{N}} A_{ij}|e)\forall A_{ij} \in \mathcal{Y} \tag{4.60}$$

Now consider any $x := (x_i)_{i\in\mathbb{N}} \in X^{\mathbb{N}}$. By definition of e, $e_{x_i i} = x_i$ for any $i, j \in \mathbb{N}$.

Define

$$\mathbb{Q} : X^{\mathbb{N}} \rightarrowtriangle Y^{\mathbb{N}} \tag{4.61}$$

$\mathbb{Q} :=$ 

(4.62)

and consider some $A \subset \mathbb{N}$, $|A| = n$ and $B := (x_i, i))_{i\in A}$. Note that the subsequence of e indexed by B, $e_B := (e_{x_i i})_{i\in A} = x_A$. Thus given the swap map $\text{swap}_{A\leftrightarrow B} : \mathbb{N} \to \mathbb{N}$ that sends the first element of A to the first element of B and so forth, $\text{swap}_{A\leftrightarrow B}(e_B) = x_A$. For arbitrary $\{C_i \in \mathcal{Y}|i \in A\}$, define $C_A := \text{swap}_{[n]\leftrightarrow A}(\times_{i\in[n]} C_i \times Y^{\mathbb{N}})$. Then, for arbitrary $x \in X^{\mathbb{N}}$

$$\mathbb{Q}(C_A|x) = \mu(\text{ev}_x^{-1}(C_A)) \tag{4.63}$$

The argument of μ is

$$\text{ev}_x^{-1}(C_A) = \{(y_{ji})_{j,i\in X\times\mathbb{N}}|(y_{x_i i})_{i\in\mathbb{N}} \in C_A\} \tag{4.64}$$

$$= \bigtimes_{i\in\mathbb{N}} \bigtimes_{j\in X} D_{ji} \tag{4.65}$$

where

$$D_{ji} = \begin{cases} C_i & (j, i) \in B \\ Y & \text{otherwise} \end{cases} \tag{4.66}$$

and so

$$\text{swap}_{A\leftrightarrow B}(\text{ev}_x^{-1}(C_A)) = C_A \tag{4.67}$$

Substituting Equation 4.67 into 4.63

$$\begin{aligned}
\mathbb{Q}(C_A|x) &= \mu\mathrm{swap}_{A\leftrightarrow B}(C_A) && (4.68)\\
&= \mathbb{K}\mathrm{swap}_{A\leftrightarrow B}(C_A|e) && (4.69)\\
&= \mathbb{K}\mathrm{swap}_{A\leftrightarrow B}(C_A|e_B, \mathrm{swap}_{B\leftrightarrow A}(x)_B^C) && \text{by locality } (4.70)\\
&= \mathbb{K}\mathrm{swap}_{A\leftrightarrow B}(C_A|\mathrm{swap}_{B\leftrightarrow A}(x)) && (4.71)\\
&= \mathrm{swap}_{B\leftrightarrow A}\mathbb{K}\mathrm{swap}_{A\leftrightarrow B}(C_A|x) && (4.72)\\
&= \mathbb{K}(C_A|x) && \text{by commutativity of exchange } (4.73)
\end{aligned}$$

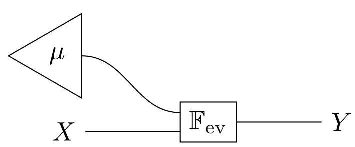
Because this holds for all x, $A \subset \mathbb{N}$, by Lemma 4.3.5

$$\mathbb{Q} = \mathbb{K} \tag{4.74}$$

Next we will show μ is column exchangeable. Consider any column swap $\mathrm{swap}_c : X \times \mathbb{N} \to X \times \mathbb{N}$ that acts as the identity on the X component and a finite permutation on the $\mathbb{N}$ component. From the definition of e, $\mathrm{swap}_c(e) = e$. Thus by commutativity of exchange, for any $A \in \mathcal{Y}^{\mathbb{N}}$

$$\begin{aligned}
\mathbb{K}(A|e) &= \mathrm{swap}_c\mathbb{K}\mathrm{swap}_c(A|e) && (4.75)\\
&= \mathbb{K}\mathrm{swap}_c(A|\mathrm{swap}_c(e)) && (4.76)\\
&= \mathbb{K}\mathrm{swap}_c(A|e) && (4.77)
\end{aligned}$$

If: Suppose

$\mathbb{K} =$  (4.78)

where μ is column exchangeable, and consider any two $x, x' \in X^{\mathbb{N}}$ such that some subsequences are equal $x_S = x'_T$ with $S, T \subset \mathbb{N}$ and $|S| = |T| = [n]$.

For any $\{A_i \in \mathcal{Y}|i \in S\}$, let $A_S = \mathrm{swap}_{[n]\leftrightarrow S} \times_{i\in[n]} A_i \times Y^{\mathbb{N}}$, $A_T = \mathrm{swap}_{S\leftrightarrow T}(A_S)$, $B = (x_i i)_{i\in S}$ and $C = (x_i i)_{i\in T} = (x_{\mathrm{swap}_{S\leftrightarrow T}}(i)i)_{i\in S}$. By Equations 4.63 and 4.67

$$\begin{aligned}
\mathbb{K}(A_S|x) &= \mu\mathrm{swap}_{S\leftrightarrow B}(A_S) && (4.79)\\
&= \mu\mathrm{swap}_{T\leftrightarrow C}(A_T) && \text{by column exchangeability of } \mu \ (4.80)\\
&= \mathbb{K}(A_T|\mathrm{swap}_{S\leftrightarrow T}(x)) && (4.81)\\
&= \mathrm{swap}_{S\leftrightarrow T}\mathbb{K}(A_T|x) && (4.82)\\
&= \mathrm{swap}_{S\leftrightarrow T}\mathbb{K}\mathrm{swap}_{S\leftrightarrow T}(A_S|x) && (4.83)
\end{aligned}$$

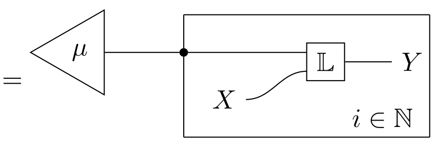
so $\mathbb{K}$ is causally contractible by Theorem 4.3.7. $\square$

Theorem 4.3.11 is the main result of this section. It shows that a causally contractible Markov kernel $X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ is representable as a "prior" $\mu \in \Delta(H)$ and a "parallel product" of Markov kernels $H \times X \rightarrow Y$. These will be the response conditionals when Theorem 4.3.11 is applied to probability set models.

Definition 4.3.10 (Measurable set of probability distributions)**.** Given a measurable set $(\Omega, \mathcal{F})$, the measurable set of distributions on Ω, $\mathcal{M}_1(\Omega)$, is the set of all probability distributions on Ω equipped with the coarsest σ-algebra such that the evaluation maps $\eta_B : \nu \mapsto \nu(B)$ are measurable for all $B \in \mathcal{F}$.

Theorem 4.3.11. *Given a kernel* $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$*, let* $(H, \mathcal{H}) := \mathcal{M}_1(Y^X)$ *be the set of probability distributions on* $(Y^X, \mathcal{Y}^X)$*.* $\mathbb{K}$ *is causally contractible if and only if there is some* $\mu \in \Delta(H)$ *and* $\mathbb{L} : H \times X \rightarrow Y$ *such that*

$$\mathbb{K} =$$



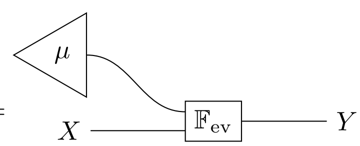
(4.84)

$$\iff \tag{4.85}$$

$$\mathbb{K}(\bigtimes_{i\in\mathbb{N}} A_i|(x_i)_{i\in\mathbb{N}}) = \int_H \prod_{i\in\mathbb{N}} \mathbb{L}(A_i|h, x_i)\mu(\mathrm{d}h) \tag{4.86}$$

Proof. By Theorem 4.3.9, we can represent the conditional probability $\mathbb{K}$ as

$$\mathbb{K} =$$

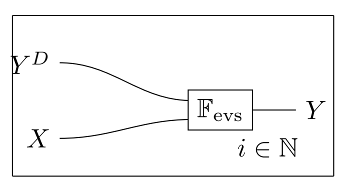


(4.87)

where μ is column exchangeable.

As a preliminary, we will show

$$\mathbb{F}_{\mathrm{ev}} =$$



(4.88)

where $\mathrm{evs}_{Y^D \times D} : Y^D \times D \rightarrow Y$ is the single-shot evaluation function

$$(x, (y_i)_{i\in X}) \mapsto y_x \tag{4.89}$$

Recall that ev is the function

$$((x_i)_{\mathbb{N}}, (y_{ji})_{j,i\in X\times\mathbb{N}}) \mapsto (y_{x_i i})_{i\in\mathbb{N}} \tag{4.90}$$

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

$$\mathbb{F}_{\text{ev}}(\bigtimes_{i\in\mathbb{N}} A_i|(x_i)_{\mathbb{N}}, (y_{ji})_{i\in X\times\mathbb{N}}) = \delta_{(y_{x_i i})_{i\in\mathbb{N}}}(\bigtimes_{i\in\mathbb{N}} A_i) \tag{4.91}$$

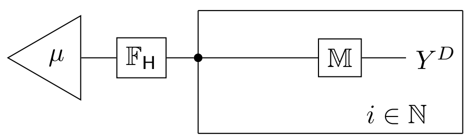
$$= \prod_{i\in\mathbb{N}} \delta_{y_{x_i i}}(A_i) \tag{4.92}$$

$$= \prod_{i\in\mathbb{N}} \mathbb{F}_{\text{evs}}(A_i|x_i, (y_{ji})_{j\in X}) \tag{4.93}$$

$$= \left(\bigotimes_{i\in\mathbb{N}} \mathbb{F}_{\text{evs}}\right)(\bigtimes_{i\in\mathbb{N}} A_i|(x_i)_{\mathbb{N}}, (y_{ji})_{j\in X\times\mathbb{N}}) \tag{4.94}$$

which is what we wanted to show.

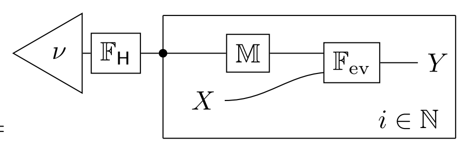
Only if: Define $\mathbb{M} : H \to Y^D$ by $\mathbb{M}(A|h) = h(A)$ for all $A \in \mathcal{Y}^X$, $h \in H$. By the column exchangeability of μ, from Kallenberg (2005a, Prop. 1.4) there is a directing random measure $\mathsf{H} : Y^{X\times\mathbb{N}} \to H$ such that

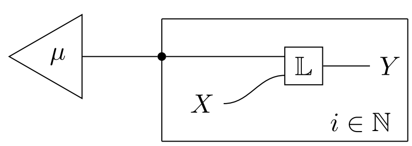
$\mu =$  (4.95)

$$\iff \tag{4.96}$$

$$\mu(\bigtimes_{i\in\mathbb{N}} A_i) = \int_H \prod_{i\in\mathbb{N}} \mathbb{M}(A_i|h)\mu\mathbb{F}_{\mathsf{H}}(\mathrm{d}h) \qquad \forall A_i \in \mathcal{Y}^X \tag{4.97}$$

By Equations 4.87 and 4.88

$\mathbb{K} =$  (4.98)

$:=$  (4.99)

Where we can connect the copied outputs of $\mu\mathbb{F}_{\mathsf{H}}$ to the inputs of each $\mathbb{M}$ "inside the plate" as the plates in Equations 4.88 and 4.95 are equal in number and each connected wire represents a single copy of Y^D.

If: By assumption, for any $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$, $x := (x_i)_{i\in\mathbb{N}} \in X^{\mathbb{N}}$

$$\mathbb{K}(\bigtimes_{i\in\mathbb{N}} A_i|x) = \int_H \prod_{i\in\mathbb{N}} \mathbb{L}(A_i|h, x_i)\mu(\mathrm{d}h) \tag{4.100}$$

Consider any $S, T \subset \mathbb{N}$ with $|S| = |T|$, and define $A_S := \times_{i\in\mathbb{N}} B_i$ where $B_i = Y$ if $i \notin S$, otherwise A_i is an arbitrary element of $\mathcal{Y}$. Define $A_T := \times_{i\in\mathbb{N}} B_{\text{swap}_{S\leftrightarrow T}(i)}$.

$$\mathbb{K}(A_S|x) = \int_H \prod_{i\in S} \mathbb{L}(A_i|h, x_i)\mu(\mathrm{d}h) \tag{4.101}$$

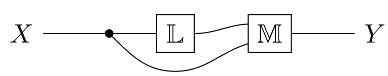
$$= \int_H \prod_{i\in T} \mathbb{L}(A_i|h, x_{\mathrm{swap}_{S\leftrightarrow T}(i)})\mu(\mathrm{d}h) \tag{4.102}$$

$$= \mathrm{swap}_{S\leftrightarrow T}\mathbb{K}(A_T|x) \tag{4.103}$$

$$= \mathrm{swap}_{S\leftrightarrow T}\mathbb{K}\mathrm{swap}_{S\leftrightarrow T}(A_S|x) \tag{4.104}$$

So by Theorem 4.3.7, $\mathbb{K}$ is causally contractible. $\square$

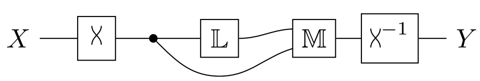
Lemma 4.3.12. *A kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrowtriangle Y^{\mathbb{N}}$ with X, Y standard measurable commutes with exchange if and only if there is some $\mathbb{L} : X^{\mathbb{N}} \rightarrowtriangle Z$ symmetric in its inputs and $\mathbb{M} : Z \times X \rightarrowtriangle Y$ such that $\mathbb{M}(\cdot|z, \cdot)$ is causally contractible for every $z \in Z$ and*



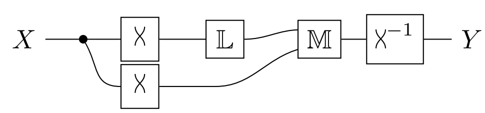
$$\mathbb{K} \overset{\mu}{\cong} \tag{4.105}$$

for any exchangeable μ.

Proof. If: Swap maps are deterministic, so by Theorem 2.3.6

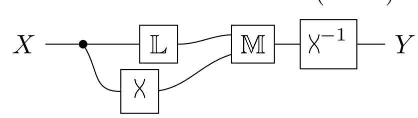


$=$



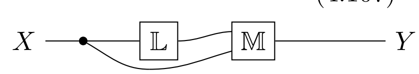
(4.106)

$=$



(4.107)

$=$



(4.108)

where Equation 4.107 follows from symmetry of $\mathbb{L}$ and 4.108 follows from causal contractibility of $\mathbb{M}(\cdot|z, \cdot)$

Only if: Construct the probability space $(\mathbb{Q}, X^{\mathbb{N}} \times Y^{\mathbb{N}}, \mathcal{X}^{\mathbb{N}} \otimes \mathcal{Y}^{\mathbb{N}})$ with $\mathbb{P} = \mu \odot \mathbb{K}$.

Take all sets in $\mathcal{X}^{\mathbb{N}}$ invariant under any finite permutation, and call this the *exchangeable σ-algebra* $\mathcal{E}$ (Kallenberg, 2005a, pg. 29). Let $\mathcal{E}_{\mathsf{X}} := \mathsf{X}^{-1}(\mathcal{E})$.

Consider $\mathbb{Q}^{\mathsf{Y}_{[n]}|\mathsf{X}}$ for some $n \in \mathbb{N}$. This is equivalent to $\mathbb{K}(\mathrm{id}_{Y^n} \otimes \mathrm{del}_{Y^{\mathbb{N}}})$. By assumption, $\mathbb{K}(\mathrm{id}_{Y^n} \otimes \mathrm{del}_{Y^{\mathbb{N}}})$ is invariant to any finite permutation of X that only affects indices after the nth. That is, $\mathbb{Q}^{\mathsf{Y}_{[n]}|\mathsf{X}}$ is $\mathcal{E}_{\mathsf{X}_{[n+1,\infty)}} \vee \sigma(\mathsf{X}_{[n]}$-measurable.

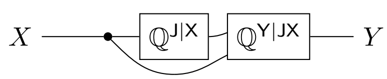
Let $\mathcal{T}_{\mathsf{X}}$ be the tail σ-algebra on $X^{\mathbb{N}}$. $\mathcal{T}_{\mathsf{X}}$ is defined as the intersection $\cap_{i=1}^{\infty}\sigma(\mathsf{X}_{[i,\infty)})$. Note that $\mathcal{T}_{\mathsf{X}_{[n+1,\infty)}} = \mathcal{T}_{\mathsf{X}}$. By Kallenberg (2005a, Corollary 1.6), $\mathcal{E}_{\mathsf{X}_{[n+1,\infty)}} = \mathcal{T}_{\mathsf{X}_{[n+1,\infty)}}$ almost surely, and so $\mathcal{E}_{\mathsf{X}_{[n,\infty)}} = \mathcal{T}_{\mathsf{X}}$ almost surely and by Kallenberg (2005a, Corollary 1.6) again $\mathcal{E}_{\mathsf{X}_{[n,\infty)}} = \mathcal{E}_{\mathsf{X}}$ μ-almost surely.

Thus $\mathbb{Q}^{\mathsf{Y}_{[n]}|\mathsf{X}}$ is $\mathcal{E}_{\mathsf{X}} \vee \sigma(\mathsf{X}_{[n]})$-measurable. By Kallenberg (2005a, Corollary 1.6) again, there is a random J taking values in the set of distributions on X such that $\sigma(\mathsf{J}) = \mathcal{E}_{\mathsf{X}}$ μ-almost surely. Thus

$$\mathbb{Q}^{\mathsf{Y}_{[n]}|\mathsf{X}\mathsf{J}} = \mathbb{Q}^{\mathsf{Y}_{[n]}|\mathsf{X}_{[n]}\mathsf{J}} \otimes \mathrm{erase}_{X^{\mathbb{N}}} \tag{4.109}$$

That is, $\mathsf{Y}_{[n]} \perp\!\!\!\perp_{\mathbb{Q}} \mathsf{X}_{[n+1,\infty)}|(\mathsf{X}_{[n]}, \mathsf{J})$. In particular, $\mathbb{Q}^{\mathsf{Y}|\mathsf{J}\mathsf{X}}(\cdot|z,\cdot)$ is local for all $z \in \Delta(X)$. By disintegration

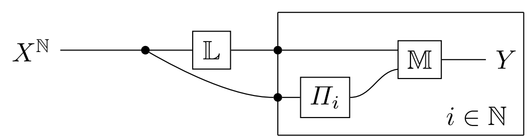
$$\mathbb{K} \overset{\mu}{\cong} \tag{4.110}$$



which completes the proof. □

Theorem 4.3.13. *Given a kernel $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$, let $(H, \mathcal{H}) := \mathcal{M}_1(Y^X)$ be the measurable set of probability distributions on $(Y^X, \mathcal{Y}^X)$. $\mathbb{K}$ is exchange commutative if and only if there is some $\mathsf{H} : Y^{X\times\mathbb{N}} \to H$, some $\mathbb{M} : X^{\mathbb{N}} \rightarrow H$ symmetric in its inputs and some $\mathbb{L} : H \times X \rightarrow Y$ such that*

$$\mathbb{K} \overset{\nu}{\cong} \tag{4.111}$$



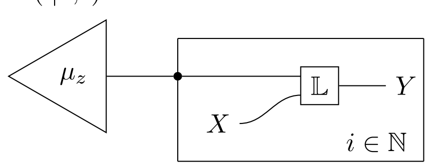
$$\Longleftrightarrow \tag{4.112}$$

$$\mathbb{K}(\bigtimes_{i\in\mathbb{N}} A_i|x) \overset{\nu}{\cong} \int_H \prod_{i\in\mathbb{N}} \mathbb{M}(A_i|h, x_i)\mathbb{L}(\mathrm{d}h|x) \tag{4.113}$$

for arbitrary exchangeable ν, where $\Pi_i : X^{\mathbb{N}} \rightarrow X$ is the Markov kernel associated with the i-th projection map.

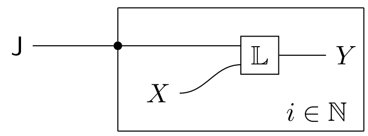
Proof. Only if: By Lemma 4.3.12, $\mathbb{K} = (\mathrm{id}_{X^{\mathbb{N}}} \odot \mathbb{L})\mathbb{M}$ where $\mathbb{M} : J \times X^{\mathbb{N}} \rightarrow Y$ is such that $\mathbb{M}(\cdot|z,\cdot)$ is causally contractible for each $z \in J$. Thus by Theorem 4.3.11, for each z we have

$$\mathbb{M}_z := \mathbb{M}(\cdot|z,\cdot) \tag{4.114}$$



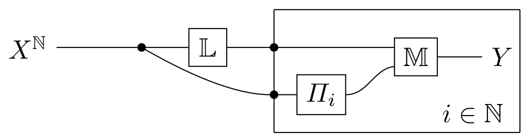
$$= \tag{4.115}$$

Where, in particular, $\mathbb{L}$ is the same for all z. Thus

$\mathbb{M} =$ 

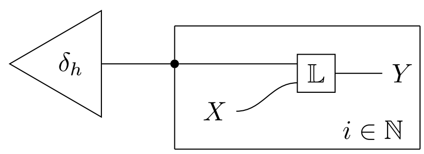
(4.116)

Hence

$\mathbb{K} \overset{\nu}{\cong}$ 

(4.117)

If: Apply Lemma 4.3.12 identifying $\mathbb{M}$ with $\mathbb{L}$. By Theorem 4.3.11,



(4.118)

is causally contractible for every $h \in H$. □

4.3.3 Conditionally independent and identical response functions with data-independent inputs

In this section, we apply the theorems from the previous part to models $\mathbb{P}_C$ with "input" variables $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and corresponding "output" variables $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$ where the inputs are independent of non-corresponding output variables. Data-independent models are interesting because they are precisely the class of models with a uniform conditional distribution $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ that is causally contractible. This makes them simple to work with in comparison with more general models. The assumption of data independence is a very limiting assumption – see-do models, for example (Definition 3.2.13) do not satisfy it. More general cases are considered in Sections 4.4 and 4.5.

Call a model $\mathbb{P}_C$ with sequential outputs Y and a corresponding sequence of inputs D a "sequential input-output model".

Definition 4.3.14 (Sequential input-output model)**.** A *sequential input-output model* is a triple $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ where $\mathbb{P}_C$ is a probability set on $(\Omega, \mathcal{F})$, D is a sequence of "inputs" $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and Y is a corresponding sequence of "outputs" $\mathsf{Y} = (\mathsf{Y}_i)_{i\in\mathbb{N}}$ where $\mathsf{D}_i : \Omega \to D$ and $\mathsf{Y}_i : \Omega \to Y$.

Given a sequence of variables $(\mathsf{D}_i, \mathsf{Y}_i)_{i\in\mathbb{N}}$ where the "inputs" are $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and the "outputs" are $\mathsf{Y} = (\mathsf{Y}_i)_{i\in\mathbb{N}}$, say the inputs are data-independent if $\mathsf{D}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}_{[i-1]}|(\mathsf{D}_{[i-1]}, \mathsf{C})$ for all $i \in \mathbb{N}$. This could model an experiment where it may be possible to choose different inputs D, but all the inputs are determined before the outputs Y are known.

Theorem 4.3.15 applies Theorem 4.3.11 to the case of a sequential input-output model where $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ is causally contractible. It shows that these are precisely the models $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ with data-independent inputs, conditionally independent and identical response functions $\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}$ and data-independent hypotheses.

The a requirement that all values of D appear infinitely often allows us to define the hypothesis H as a function of the observed variables. Without it, it's possible to define a latent H, but it is not generally a function of observed variables.

Theorem 4.3.15 (Data-independent causal contractibility)**.** *Given a sequential input-output model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *on* $(\Omega, \mathcal{F})$ *with D countable, suppose* $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{C}|\mathsf{D}$ *and, letting $E \subset D^{\mathbb{N}}$ be the set of all sequences d such that each $x \in D$ appears in d infinitely often, suppose also that $\mathbb{P}_\alpha^{\mathsf{D}}(E) = 1$ for all $\alpha \in C$. Then* $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ *is causally contractible if and only if there is some directing random measure* $\mathsf{H} : \Omega \to H$ *such that* $\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}} = \mathbb{P}_C^{\mathsf{Y}_j|\mathsf{D}_j\mathsf{H}}$ *for all $i, j \in \mathbb{N}$,* $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{[i-1]}, \mathsf{D}_{[i-1}, \mathsf{C})|(\mathsf{D}_i, \mathsf{H})$*,* $\mathsf{D}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}_{[i-1]}|(\mathsf{D}_{[i-1]}, \mathsf{C})$ *and* $\mathsf{H} \perp\!\!\!\perp_{\mathbb{P}_C} (\mathsf{D}, \mathsf{C})$*.*

Proof. If: First, construct the variable Y^D. Let $\#^i_{\mathsf{D}_k=j}$ be 1 if $\mathsf{D}_k = j$ and it is exactly the ith coordinate of D that does so, and 0 otherwise. Then for $i \in \mathbb{N}$ and $j \in D$

$$\mathsf{Y}^D_{ij} = \sum_{k=1}^{\infty} \#^i_{\mathsf{D}_k=j} \mathsf{Y}_k \tag{4.119}$$

That is, the (i, j)-th coordinate of Y^D is equal to the coordinate Y_k for which the corresponding $\mathsf{D}_k = j$ and there have been $i - 1$ preceding instances where a coordinate of D has been equal to j. Note that Y^D_{ij} is equal to some Y_k with probability 1, by the assumption that all values of D occur infinitely often almost surely.

Note that $\#^i_{\mathsf{D}_k=j}$ and $\#^i_{\mathsf{D}_l=j}$ are mutually exclusive for $k \neq l$. Thus there is a random permutation R taking values in the set of functions $\mathbb{N} \to \mathbb{N}$ (including infinite permutations) such that

$$\mathsf{Y}^D = (\mathsf{Y}_{\mathrm{ef}(\mathsf{R}, i)}) := \mathsf{Y}_{\mathsf{R}} \tag{4.120}$$

where ef is the function evaluation map $\mathrm{ef}(f, x) = f(x)$. By construction, $\mathsf{D}_{\mathsf{R}} \bmod |D| = d_{i \bmod |D|}$ i.e. the $i \bmod |D|$'th element of D.

Now,

$$\mathbb{P}_\alpha^{\mathsf{Y}_{\mathsf{R}}|\mathsf{D}_{\mathsf{R}}}(A|d) = \int_R \mathbb{P}_\alpha^{\mathsf{Y}_\rho|\mathsf{D}_\rho}(A|d) \mathbb{P}_\alpha^{\mathsf{R}}(mathrmd\rho) \tag{4.121}$$

$$\tag{4.122}$$

For each ρ, define $\rho^n : \mathbb{N} \to \mathbb{N}$ as the permutation that agrees with ρ on the first n indices and is the identity otherwise. By causal contractibility, for $n \in \mathbb{N}$

$$\mathbb{P}^{\mathsf{Y}_{\rho^n([n])}|\mathsf{D}_{\rho^n([n])}} = \mathbb{P}^{\mathsf{Y}_{\rho([n])}|\mathsf{D}_{\rho([n])}} \tag{4.123}$$

$$= \mathbb{P}^{\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}} \tag{4.124}$$

By Lemma 4.3.5, it must therefore be the case that

$$\mathbb{P}^{\mathsf{Y}|\mathsf{D}} = \mathbb{P}^{\mathsf{Y}_\rho|\mathsf{D}_\rho} \tag{4.125}$$

Then from Equation 4.121

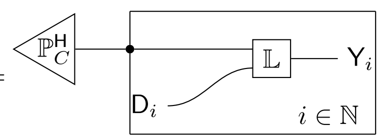
$$\mathbb{P}_\alpha^{\mathsf{Y}_\mathsf{R}|\mathsf{D}_\mathsf{R}}(A|d) = \int_R \mathbb{P}_\alpha^{\mathsf{Y}_\rho|\mathsf{D}_\rho}(A|d)\mathbb{P}_\alpha^{\mathsf{R}}(mathrmd\rho) \tag{4.126}$$

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

$$\mathbb{P}_\alpha^{\mathsf{Y}^D}(A) = \mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}(A|e) \tag{4.127}$$

By exchange commutativity and invariance of e_i to permutations of period $|D|$, $\mathbb{P}_\alpha^{\mathsf{Y}^D}$ is column exchangeable. Thus from Kallenberg (2005a, Prop. 1.4), there exists a directing random measure $\mathsf{H} : Y^{D\times\mathbb{N}} \to H$ such that

$\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}} =$ (4.128)



it remains to be shown that $\mathbb{L}$ is a version of $\mathbb{P}^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}$ for all $i \in \mathbb{N}$, $\mathsf{D}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}_{[i-1]}|(\mathsf{D}_{[i-1]}, \mathsf{C})$ and $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{[i-1]}, \mathsf{D}_{[i-1]}, C)|(\mathsf{D}_i, \mathsf{H})$.

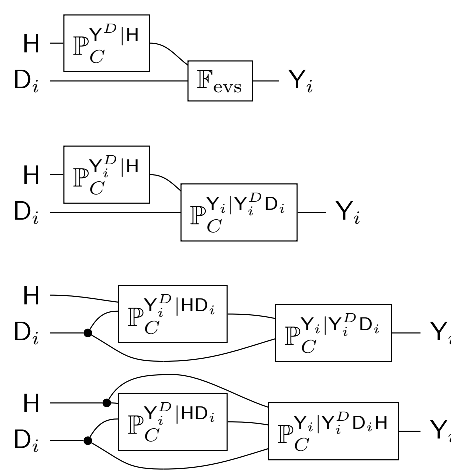
To show $\mathbb{L}$ is a version of $\mathbb{P}^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i}$ for all $i \in \mathbb{N}$:

$\mathbb{L} =$ (4.129)

$=$ (4.130)

$=$ (4.131)

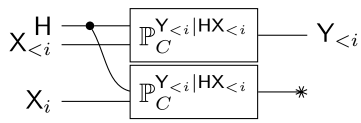
$=$ (4.132)

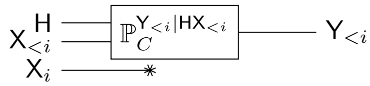


$$= \mathbb{P}_C^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i} \tag{4.133}$$

Where 4.131 follows from $\mathsf{H} \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{D}_i, \mathsf{C})$, which itself follows from $\mathsf{Y}_i^D \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{D}_i, \mathsf{C})$ which holds by construction. 4.132 follows from $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{H}, \mathsf{C})|(\mathsf{Y}_i^D, \mathsf{D}_i)$, which follows from Y_i being a deterministic function of $(\mathsf{Y}_i^D, \mathsf{D}_i)$.

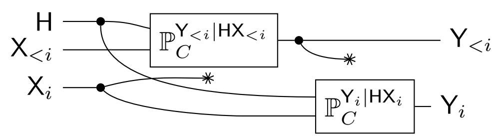
For the independence $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{[i-1]}, \mathsf{D}_{[i-1]}, C)|(\mathsf{D}_i, \mathsf{H})$, note that

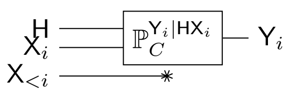
$\mathbb{P}_C^{\mathsf{Y}_{<i}|\mathsf{H}\mathsf{X}_{<i}\mathsf{X}_i} =$  (4.134)

$=$  (4.135)

hence $\mathsf{Y}_{<i} \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{X}_i, \mathsf{C})|(\mathsf{H}, \mathsf{X}_{<i})$

Then

$\mathbb{P}_C^{\mathsf{Y}_i\mathsf{Y}_{<i}|\mathsf{H}\mathsf{D}_i\mathsf{D}_{<i}} =$  (4.136)

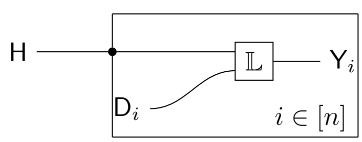
$\implies \mathbb{P}_C^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i\mathsf{D}_{<i}} \overset{\mathbb{P}_C}{\cong}$  (4.137)

by Theorem 2.4.33. Hence $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{X}_{<i}, \mathsf{Y}_{<i}, \mathsf{C})|(\mathsf{H}, \mathsf{X}_i)$.

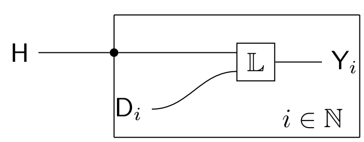
Only if: By assumption, for all $i \in \mathbb{N}$

$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{H}\mathsf{X}_{[i]}\mathsf{Y}_{<i}} \overset{\mathbb{P}_C}{\cong} \mathrm{del}_{X^{i-1}\times Y^{i-1}} \otimes \mathbb{P}_C^{\mathsf{Y}_1|\mathsf{H}\mathsf{X}_1} \tag{4.138}$$

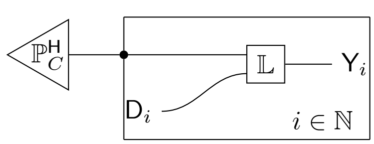
Thus for all $n \in \mathbb{N}$ by repeated application of Theorem 2.4.33

$\mathbb{P}_C^{\mathsf{Y}_{[n]}|\mathsf{H}\mathsf{X}_{[n]}} \overset{\mathbb{P}_C}{\cong}$  (4.139)

thus by Lemma 4.3.5

$\mathbb{P}_C^{\mathsf{Y}_{\mathbb{N}}|\mathsf{H}\mathsf{X}_{\mathbb{N}}} \overset{\mathbb{P}_C}{\cong}$  (4.140)

and, because $\mathsf{H} \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{X}, \mathsf{C})$

$\mathbb{P}_C^{\mathsf{Y}_{\mathbb{N}}|\mathsf{X}_{\mathbb{N}}} \overset{\mathbb{P}_C}{\cong}$  (4.141)

causal contractibility follows from Theorem 4.3.11. □

A consequence of Theorem 4.3.7 applied to just-do models $\mathbb{P}_C$ with causally contractible $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ is that, for any $A, B \subset \mathbb{N}$ with $|A| = |B|$, $\mathbb{P}_C^{\mathsf{Y}_A|\mathsf{D}_A} = \mathbb{P}_C^{\mathsf{Y}_B|\mathsf{D}_B}$. A further consequence is the interchangeability of conditioning data – for any $i \in \mathbb{N}$, $\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{Y}_A\mathsf{D}_A} = \mathbb{P}_C^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{Y}_B\mathsf{D}_B}$.

Theorem 4.3.16 (Equality of subsequence conditionals)**.** *A sequential just-do model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *with* $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ *causally contractible satisfies, for any* $A, B \subset \mathbb{N}$ *with* $|A| = |B|$

$$\mathbb{P}_C^{\mathsf{Y}_A|\mathsf{D}_A} \overset{\mathbb{P}_C}{\cong} \mathbb{P}_C^{\mathsf{Y}_B|\mathsf{D}_B} \tag{4.142}$$

Proof. Only if: For any $A, B \subset \mathbb{N}$, let $\mathrm{swap}_{B\leftrightarrow A,D} : D^{\mathbb{N}} \rightarrowtriangle D^{\mathbb{N}}$ be the transpositon of B with A indices and $\mathrm{swap}_{B\leftrightarrow A,Y} : Y^{\mathbb{N}} \rightarrowtriangle Y^{\mathbb{N}}$ be the same defined on Y. By Theorem 4.3.7

$$\mathbb{P}_C^{\mathsf{Y}_A|\mathsf{D}_A} \otimes \mathrm{del}_{D^{\mathbb{N}}} = \mathbb{P}_C^{\mathsf{Y}|\mathsf{D}} \mathrm{marg}_A \tag{4.143}$$

$$= \mathrm{swap}_{A\leftrightarrow[n],D} \mathbb{P}_C^{\mathsf{Y}_{[n]}|\mathsf{D}} \tag{4.144}$$

$$= \mathbb{P}_C^{\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}} \otimes \mathrm{del}_{D^{\mathbb{N}}} \tag{4.145}$$

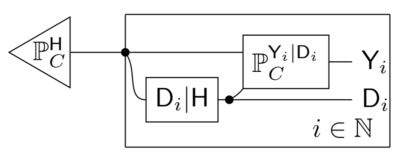
$$= \mathrm{swap}_{N\leftrightarrow[n],D} \mathbb{P}_C^{\mathsf{Y}_{[n]}|\mathsf{D}} \tag{4.146}$$

$$= \mathbb{P}_C^{\mathsf{Y}_B|\mathsf{D}_B} \otimes \mathrm{del}_{D^{\mathbb{N}}} \tag{4.147}$$

□

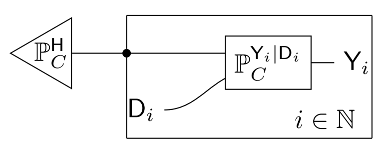
Examples

Purely passive observations can be modeled with a probability set $\mathbb{P}_C$ where $|\mathbb{P}_C| = 1$. In this case, a model that is exchangeable over the sequence of pairs $(\mathsf{D}_i, \mathsf{Y}_i)_{i\in\mathbb{N}}$ has $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ causally contractible. This follows from the fact that

$\mathbb{P}_C^{\mathsf{YD}} =$ 

(4.148)

and so



(4.149)

is a version of $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$.

Instead of passive observations only, a model might feature a subsequence of passive observations and a subsequence of active interventions. Say the passive observations are $(\mathsf{D},\mathsf{Y})_{i\in\mathbb{N}}$ and the active interventions are $(\mathsf{E},\mathsf{Z})_{i\in\mathbb{N}}$. By the previous argument, $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ is causally contractible. We might further assume that $\mathbb{P}_C^{\mathsf{YZ}|\mathsf{DE}}$ is causally contractible – that is, there is a independent and identical response function $\mathbb{P}_C^{\mathsf{Z}_i|\mathsf{E}_i\mathsf{H}}$ equal to $\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}$.

One consequence of this is "observational imitation": any choice α that makes $\mathbb{P}_\alpha^{\mathsf{DE}}$ exchangeable also makes $\mathbb{P}_\alpha^{\mathsf{YZ}}$ exchangeable. That is, if for some permutation swap_ρ

$$\mathbb{P}_\alpha^{\mathsf{DE}}\text{swap}_\rho = \mathbb{P}_\alpha^{\mathsf{DE}} \tag{4.150}$$

then by commutativity of exchange

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{YZ}} &= \mathbb{P}_\alpha^{\mathsf{DE}}\mathbb{P}_C^{\mathsf{YZ}|\mathsf{DE}} && (4.151)\\
&= \mathbb{P}_\alpha^{\mathsf{DE}}\text{swap}_\rho\mathbb{P}_C^{\mathsf{YZ}|\mathsf{DE}} && (4.152)\\
&= \mathbb{P}_\alpha^{\mathsf{DE}}\mathbb{P}_C^{\mathsf{YZ}|\mathsf{DE}}\text{swap}_\rho && (4.153)\\
&= \mathbb{P}_C^{\mathsf{YZ}|\mathsf{DE}}\text{swap}_\rho && (4.154)
\end{aligned}$$

However, the assumption that $\mathbb{P}_C^{\mathsf{YZ}|\mathsf{DE}}$ is causally contractible seems unreasonable in most situations. One implication of this assumption is (by Theorem 4.3.7):

$$\begin{aligned}
&\mathbb{P}_C^{\mathsf{YZ}_i|\mathsf{DE}_i} = \mathbb{P}^{\mathsf{Z}|\mathsf{E}} && (4.155)\\
\implies &\mathbb{P}_C^{\mathsf{Z}_i|\mathsf{E}_i\mathsf{DY}} = \mathbb{P}^{\mathsf{Z}_i|\mathsf{E}_i\mathsf{E}_{\{i\}^C}\mathsf{Z}_{\{i\}^C}} && (4.156)
\end{aligned}$$

That is, the model must yield the same result when conditioned on either the observational results, or the results of other active interventions. It is rare to assume *a priori* that observational and experimental data are equally informative. Such a conclusion could be drawn *after* reviewing both sequences of data, see for example Eckles and Bakshy (2021), or it might be rejected Gordon et al. (2018, 2022).

Example 4.3.17 (Backdoor adjustment)**.** If a sequential just-do model $(\mathbb{P}_C, (\mathsf{D},\mathsf{X}), \mathsf{Y})$ has $\mathbb{P}_C^{\mathsf{Y}|\mathsf{DX}}$ causally contractible as well as:

- $\mathsf{X}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{D}_i C|\mathsf{H}$ (X_i is extended independent of D_i conditional on H)
- $\mathbb{P}_C^{\mathsf{X}_i|\mathsf{H}} \cong \mathbb{P}_C^{\mathsf{X}_1|\mathsf{H}}$ (the distribution of X is exchangeable)

Then the model exhibits a kind of "backdoor adjustment" Pearl (2009, Chap. 1). Specifically

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}(A|d,h) &= \int_X \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{D}_i\mathsf{H}}(A|d,x,h)\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{D}_i\mathsf{H}}(\mathrm{d}x|d,h) && (4.157)\\
&= \int_X \mathbb{P}_C^{\mathsf{Y}_1|\mathsf{X}_1\mathsf{D}_1\mathsf{H}}(A|d,x,h)\mathbb{P}_C^{\mathsf{X}_i|\mathsf{H}}(\mathrm{d}x|h) && (4.158)\\
&= \int_X \mathbb{P}_C^{\mathsf{Y}_1|\mathsf{X}_1\mathsf{D}_1\mathsf{H}}(A|d,x,h)\mathbb{P}_C^{\mathsf{X}_1|\mathsf{H}}(\mathrm{d}x|h) && (4.159)
\end{aligned}$$

Equation 4.159 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.

Example: body mass index

Given a sequential just-do model $(\mathbb{P}_C, (\mathsf{B}, \mathsf{I}), \mathsf{Y})$ with $\mathsf{B} := (\mathsf{B}_i)_{i\in M}$ representing body mass index of individual I_i and $\mathsf{Y} := (\mathsf{Y}_i)_{i\in M}$ representing health outcomes of interest for the same individual, Hernán and Taubman (2008) noted that there are multiple different choices that can influence an individual's body mass index B_i in the same way. Thus $\mathsf{YI} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{C}|\mathsf{B}$ might generally be rejected, and so there may be no uniform conditional $\mathbb{P}_C^{\mathsf{Y}|\mathsf{IB}}$. In this case, $\mathbb{P}_C^{\mathsf{Y}|\mathsf{IB}}$ cannot be causally contractible because it doesn't exist.

Suppose instead a model $(\mathbb{P}_C, (\mathsf{D}, \mathsf{I}), (\mathsf{B}, \mathsf{Y}))$ is given, with $\mathsf{D} = (\mathsf{D}_i)_{i\in M}$ representing "decisions", appropriately fine-grained to satisfy

$$\mathsf{YBI} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{C}|\mathsf{D} \tag{4.160}$$
$$\mathsf{YBI} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{D}|\mathsf{C} \tag{4.161}$$

and $\mathbb{P}_C^{\mathsf{YB}|\mathsf{ID}}$ causally contractible. Then by Theorem 5.1.9 $\mathbb{P}_C^{\mathsf{Y}|\mathsf{BD}}$ is also causally contractible. In general, there may be some $U \subset H$ such that for any $h \in U$

$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{B}_i\mathsf{D}_i\mathsf{H}}(y|b,d,h) = \mathbb{P}_C^{\mathsf{Y}_i|\mathsf{B}_i\mathsf{H}}(y|b,h) \tag{4.162}$$

then, *conditioning on* $\mathsf{H} \in U$, the resulting $\mathbb{P}_{C,\mathsf{H}\in U}^{\mathsf{Y}|\mathsf{B}}$ is causally contractible.

Defining conditioning

So it may be possible to derive the fact that there is a independent and identical response conditional $\mathbb{P}_{C,\mathsf{H}\in U}^{\mathsf{Y}_i|\mathsf{HB}_i}$ if $\mathsf{H} \in U$ is implied by available data, even if it is not assumed outright.

4.4 Conditionally independent and identical response functions with varying inputs

4.5 Conditionally independent and identical response functions with data-dependent inputs

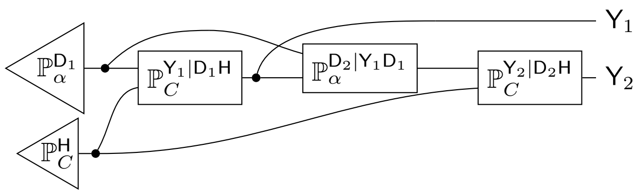
The results of the previous section concern "just-do" models where actions have not dependence on previous data. Decision problems of interest actually have actions that depend on data – what's really wanted are "see-do" models of some variety (see Definition 3.2.13). Here, Theorem 4.3.15 is generalised to sequential see-do models with the use of *probability combs*.

To begin with an example, consider a probability set $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ with $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$ as usual, and take a subsequence $(\mathsf{D}_i, \mathsf{Y}_i)_{i\in[2]}$ of length

2. Suppose $\mathbb{P}_C$ features independent and identical response conditionals in the sense that the following holds

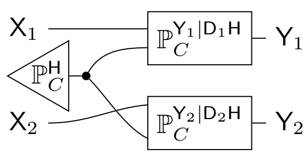
$$
\begin{aligned}
\mathsf{Y}_i &\perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{<i}, \mathsf{D}_{<i}, \mathsf{C})|\mathsf{HD}_i && \forall i \in \mathbb{N} && (4.163)\\
\wedge \mathsf{H} &\perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{D}C && && (4.164)\\
\wedge \mathbb{P}_C^{\mathsf{Y}_i|\mathsf{HD}_i} &= \mathbb{P}_C^{\mathsf{Y}_0|\mathsf{HD}_0} && \forall i \in \mathbb{N} && (4.165)
\end{aligned}
$$

Then, for arbitrary $\alpha \in C$

$\mathbb{P}_\alpha^{\mathsf{Y}_{[2]}} =$  (4.166)

note that D_2 depends on Y_1 and D_1. Instead of multiplying by a distribution over $(\mathsf{D}_1, \mathsf{D}_2)$, $\mathbb{P}_\alpha^{\mathsf{D}_2|\mathsf{Y}_1\mathsf{D}_1}$ has been "inserted" between the response conditionals $\mathbb{P}_C^{\mathsf{Y}_1|\mathsf{D}_1\mathsf{H}}$ and $\mathbb{P}_C^{\mathsf{Y}_2|\mathsf{D}_2\mathsf{H}}$. A comb is a Markov kernel that yields a probability distribution when another Markov kernel of appropriate type is inserted in this manner.

Given $\mathbb{P}_C^{\mathsf{Y}_1|\mathsf{D}_1\mathsf{H}}$ and $\mathbb{P}_C^{\mathsf{Y}_2|\mathsf{D}_2\mathsf{H}}$, define the comb

$\mathbb{P}_C^{\mathsf{Y}_{[2]}\wr\mathsf{D}_{[2]}} :=$  (4.167)

then $\mathbb{P}_C^{\mathsf{Y}_{[2]}\wr\mathsf{D}_{[2]}}$ is causally contractible. $\mathbb{P}_C^{\mathsf{Y}_{[2]}\wr\mathsf{D}_{[2]}}$ is *not* a uniform conditional probability; in general

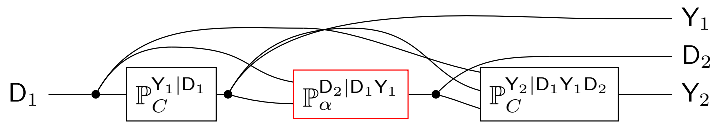
$$
\mathbb{P}_\alpha^{\mathsf{D}_1\mathsf{D}_2}\mathbb{P}_C^{\mathsf{Y}_{[2]}\wr\mathsf{D}_{[2]}} \neq \mathbb{P}_\alpha^{\mathsf{Y}_1\mathsf{Y}_2} \tag{4.168}
$$

4.5.1 Combs

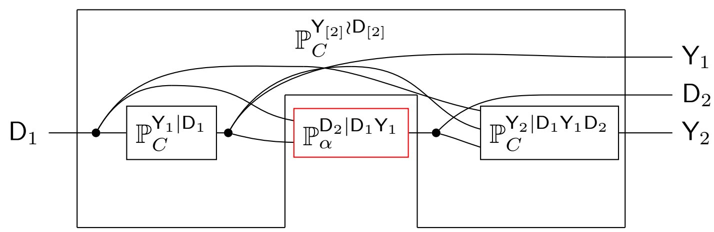
Combs generalise conditional probabilities in this sense: given a conditional distribution and a marginal distribution of the right type, joining them together (with the semidirect product2.2.22) I get a marginal distribution of a different type. Define "1-combs" as conditional probabilities and "0-combs" as conditional distributions. Then the previous observation can be restated as: given a 1-comb and a 0-comb of the right type, joining them together yields a 0-comb of a different type. Higher order combs generalise this: given an n-comb and an $n-1$-comb of the right type, joining them yields an $n-1$ comb.

Joining combs uses an “insert” operation (Definition 4.5.4). A graphical depiction of this operation gives some intuition for why it is called “insert”:

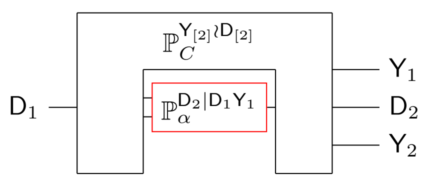
$$\mathbb{P}_\alpha^{\mathsf{Y}_1\mathsf{D}_2\mathsf{Y}_2|\mathsf{D}_1} = \text{insert}(\mathbb{P}_\alpha^{\mathsf{D}_2|\mathsf{D}_1\mathsf{Y}_1}, \mathbb{P}_C^{\mathsf{Y}_{[2]}\wr\mathsf{D}_{[2]}}) \tag{4.169}$$

$=$ 

(4.170)

$=$ 

(4.171)

$=$ 

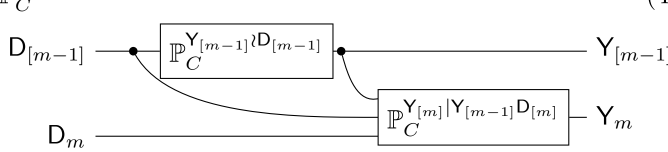
(4.172)

While Equation 4.170 is a well-formed string diagram in the category of Markov kernels, Equation 4.172 is not. In the case that all the underlying sets are discrete, Equation 4.172 can be defined using an extended string diagram notation appropriate for the category of real-valued matrices (Jacobs et al., 2019), though we do not introduce this extension here.

Formal definitions of finite and infinite combs follow, which will be used in Section 4.5.2 to generalise Theorem 4.3.15 to the data-dependent case.

Definition 4.5.1 (Uniform n-Comb)**.** Given a probability set $\mathbb{P}_C$ with variables $\mathsf{Y}_i : \Omega \to Y$, $\mathsf{D}_i : \Omega \to D$ for $i \in [n]$ and uniform conditional probabilities $\{\mathsf{P}_C^{\mathsf{Y}_i|\mathsf{D}_{[i]}\mathsf{Y}_{[i-1]}} | i \in [n]\}$, the uniform n-comb $\mathbb{P}_C^{\mathsf{Y}_{[n]}\wr\mathsf{D}_{[n]}} : D^n \rightarrow Y^n$ is the Markov kernel given by the recursive definition

$$\mathbb{P}_C^{\mathsf{Y}_1\wr\mathsf{D}_1} = \mathbb{P}_C^{\mathsf{Y}_1|\mathsf{D}_1} \tag{4.173}$$

$\mathbb{P}_C^{\mathsf{Y}_{[m]}\wr\mathsf{D}_{[m]}} =$ 

(4.174)

Definition 4.5.2 (Uniform $\mathbb{N}$-comb)**.** Given a probability set $\mathbb{P}_C$ with variables $\mathsf{Y}_i : \Omega \to Y$ and $\mathsf{D}_i : \Omega \to D$ for $i \in \mathbb{N}$ and uniform conditional probabilities $\{\mathsf{P}_C^{\mathsf{Y}_i|\mathsf{D}_{[i]}\mathsf{Y}_{[i-1]}}|i \in \mathbb{N}\}$, the uniform $\mathbb{N}$-comb $\mathbb{P}_C^{\mathsf{Y}_\mathbb{N}\wr\mathsf{D}_\mathbb{N}} : D^\mathbb{N} \rightarrow Y^\mathbb{N}$ is the Markov kernel such that for all $n \in \mathbb{N}$

$$\mathbb{P}_C^{\mathsf{Y}_\mathbb{N}\wr\mathsf{D}_\mathbb{N}}(\mathrm{id}_{Y^n} \otimes \mathrm{del}_{Y^\mathbb{N}}) = \mathbb{P}_C^{\mathsf{Y}_{[n]}\wr\mathsf{D}_{[n]}} \otimes \mathrm{del}_{Y^\mathbb{N}} \tag{4.175}$$

Theorem 4.5.3 (Existence of $\mathbb{N}$-combs)**.** *Given a probability set $\mathbb{P}_C$ with variables $\mathsf{Y}_i : \Omega \to Y$ and $\mathsf{D}_i : \Omega \to D$ for $i \in \mathbb{N}$ and uniform conditional probabilities $\{\mathsf{P}_C^{\mathsf{Y}_i|\mathsf{D}_{[i]}\mathsf{Y}_{[i-1]}}|i \in \mathbb{N}\}$, a uniform $\mathbb{N}$-comb $\mathbb{P}_C^{\mathsf{Y}_\mathbb{N}\wr\mathsf{D}_\mathbb{N}} : D^\mathbb{N} \rightarrow Y^\mathbb{N}$ exists.*

Proof. For each $n \in \mathbb{N}$ $m < n$, we have

$$\mathbb{P}_C^{\mathsf{Y}_{[n]}\wr\mathsf{D}_{[n]}}(\mathrm{id}_{Y^{n-m}}) \otimes \mathrm{del}_{Y^m}) = \mathbb{P}_C^{\mathsf{Y}_{[n-m]}\wr\mathsf{D}_{[n-m]}} \otimes \mathrm{del}_{Y^m} \tag{4.176}$$

Therefore the existence of $\mathbb{P}_C^{\mathsf{Y}_\mathbb{N}\wr\mathsf{D}_\mathbb{N}}$ is a consequence of Lemma 4.3.5. □

For discrete sets, the insert operation has a compact definition

Definition 4.5.4 (Comb insert - discrete)**.** Given an n-comb $\mathbb{P}_\alpha^{\mathsf{Y}_{[n]}\wr\mathsf{D}_{[n]}}$ and an $n-1$ comb $\mathbb{P}_\alpha^{\mathsf{D}_{[2,n]}|\mathsf{Y}_{[n-1]}}$, $(D, \mathcal{D})$ and $(Y, \mathcal{Y})$ discrete, for all $y_i \in Y$ and $d_i \in D$

$$\mathrm{insert}(\mathbb{P}_\alpha^{\mathsf{D}_{[2,n]}\wr\mathsf{Y}_{[n-1]}}, \mathbb{P}_\alpha^{\mathsf{Y}_{[n]}\wr\mathsf{D}_{[n]}})(y_{[n]}, d_{[2,n]}|d_1) = \mathbb{P}_\alpha^{\mathsf{Y}_{[n]}\wr\mathsf{D}_{[n]}}(y_n|d_{[2,n]}, d_1)\mathbb{P}_\alpha^{\mathsf{D}_{[1,n]}\wr\mathsf{Y}_{[n-1]}}(d_{[2,n]}|y_{[n-1]}) \tag{4.177}$$

4.5.2 Response conditionals in models with data dependent actions

Theorem 4.5.6 generalises Theorem 4.3.15 to models $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ with data-dependent actions, where instead the assumption that the uniform comb $\mathbb{P}_C^{\mathsf{Y}\wr\mathsf{D}}$ is causally contractible replaces the assumption that the conditional probability $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ is causally contractible.

Definition 4.5.5 (Sequential see-do model)**.** A *sequential see-do model* is a triple $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ where $\mathbb{P}_C$ is a probability set on $(\Omega, \mathcal{F})$, D is a sequence of "inputs" $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and Y is a corresponding sequence of "outputs" $\mathsf{Y} = (\mathsf{Y}_i)_{i\in\mathbb{N}}$ where $\mathsf{D}_i : \Omega \to D$ and $\mathsf{Y}_i : \Omega \to Y$ and $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} C|(\mathsf{D}_{[i]}, \mathsf{Y}_{<i})$.

Theorem 4.5.6. *Given a sequential see-do model $(\mathbb{P}'_C, \mathsf{D}', \mathsf{Y}')$ on $(\Omega, \mathcal{F})$, then $\mathbb{P}_C'^{\mathsf{Y}'\wr\mathsf{D}'}$ is causally contractible if and only if there is a latent extension $\mathbb{P}_C$ of $\mathbb{P}'_C$ to $(\Omega \times H, \mathcal{F} \otimes \mathcal{Y}^{D\times\mathbb{N}})$ with hypothesis $\mathsf{H} : \Omega \times H \to H$ such that $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}'_C} (\mathsf{Y}_{<i}, \mathsf{X}_{<i}, C)|(\mathsf{X}_i, \mathsf{H})$ and $\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{H}} = \mathbb{P}_C^{\mathsf{Y}_j|\mathsf{X}_j\mathsf{H}}$ for all $i, j \in \mathbb{N}$ and $\mathsf{H} \perp\!\!\!\perp_{\mathbb{P}_C} (\mathsf{X}, \mathsf{C})$.*

Proof. If: By assumption, there is some $\mathbb{L} : H \times D \rightarrow Y$ such that

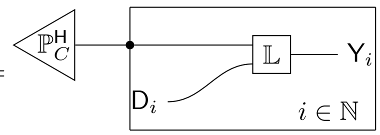
$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i} = \mathbb{L} \tag{4.178}$$

and $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{<i}, \mathsf{D}_{<i})|(\mathsf{D}_i, \mathsf{H})$. Thus

$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i\mathsf{Y}_{<i}\mathsf{D}_{<i}} = \mathbb{L} \otimes \text{erase}_{Y^{i-1}\times D^{i-1}} \tag{4.179}$$

and so

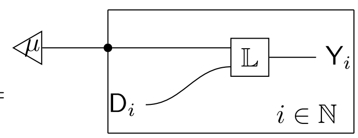
$$\mathbb{P}_C^{\mathsf{Y}\wr\mathsf{D}} = \tag{4.180}$$



and so by Theorem 4.3.11, $\mathbb{P}_C^{\mathsf{Y}\wr\mathsf{D}}$ is causally contractible.

Only if: First, define the extension $\mathbb{P}_C$. From Theorem 4.3.11 and causal contractibility of $\mathbb{P}_C^{\prime\mathsf{Y}'\wr\mathsf{D}'}$ there is some H, $\mu \in \Delta(H)$ and $\mathbb{L} : H \times D \rightarrow Y$ such that

$$\mathbb{P}_C^{\prime\mathsf{Y}'\wr\mathsf{D}'} = \tag{4.181}$$



thus, by the definition of the comb insert operation

$$\mathbb{P}_\alpha^{\prime\mathsf{D}'_{[n]}\mathsf{Y}'_{[n]}} = \mathbb{P}_\alpha^{\mathsf{D}_1} \odot \text{insert}(\mathbb{P}_\alpha^{\prime\mathsf{D}'_{[2,n]}\wr\mathsf{Y}'_{[n-1]}}, \mathbb{P}_C^{\prime\mathsf{Y}'_{[n]}\wr\mathsf{D}'_{[n]}}) \tag{4.182}$$

Let

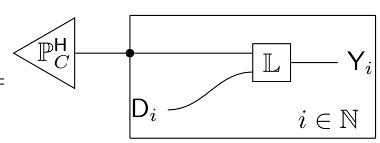
$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i} = \mathbb{L} \tag{4.183}$$

and let $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{<i}, \mathsf{D}_{<i})|(\mathsf{D}_i, \mathsf{H})$, and for all α set $\mathbb{P}_\alpha^{\mathsf{W}|\mathsf{D}\mathsf{Y}} = \mathbb{P}_\alpha^{\prime\mathsf{W}'|\mathsf{D}'\mathsf{Y}'}$ for all $\mathsf{W}' : \Omega \to W$ and $\mathbb{P}_\alpha^{\mathsf{D}_i|\mathsf{Y}_{<i}\mathsf{D}_{<i}} = \mathbb{P}_\alpha^{\prime\mathsf{D}'_i|\mathsf{Y}'_{<i}\mathsf{D}''_{<i}}$.

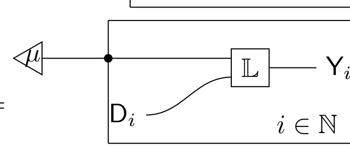
It remains to be shown that $\mathbb{P}_\alpha^{\mathsf{D}\mathsf{Y}} = \mathbb{P}_\alpha^{\prime\mathsf{D}\mathsf{Y}}$.

By Equation 4.183 and $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{<i}, \mathsf{D}_{<i})|(\mathsf{D}_i, \mathsf{H})$, it follows (for identical reasons as Equation 4.180) that

$$\mathbb{P}_C^{\mathsf{Y}\wr\mathsf{D}} = \tag{4.184}$$



$$= \tag{4.185}$$



$$= \mathbb{P}_C^{\prime\mathsf{Y}'\wr\mathsf{D}'} \tag{4.186}$$

And so for all $n \in \mathbb{N}$

$$\mathbb{P}_\alpha^{\mathsf{D}_{[n]}\mathsf{Y}_{[n]}} = \mathbb{P}_\alpha^{\mathsf{D}_1} \odot \text{insert}(\mathbb{P}_\alpha^{\mathsf{D}_{[2,n]} \wr \mathsf{Y}_{[n-1]}}, \mathbb{P}_C^{\prime \mathsf{Y}_{[n]} \wr \mathsf{D}_{[n]}}) \tag{4.187}$$

$$= \mathbb{P}_\alpha^{\mathsf{D}_1} \odot \text{insert}(\mathbb{P}_\alpha^{\prime \mathsf{D}'_{[2,n]} \wr \mathsf{Y}'_{[n-1]}}, \mathbb{P}_C^{\prime \mathsf{Y}'_{[n]} \wr \mathsf{D}'_{[n]}}) \tag{4.188}$$

$$= \mathbb{P}_\alpha^{\prime \mathsf{D}'_{[n]}\mathsf{Y}'_{[n]}} \tag{4.189}$$

□

In contrast to the data-independent case where causal contractibility of $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ implies the equivalence of all subsequence conditionals $\mathbb{P}_C^{\mathsf{Y}_A|\mathsf{X}_A}$ for all equally sized $A \subset \mathbb{N}$, a causally contractible comb $\mathbb{P}_C^{\mathsf{Y}\wr\mathsf{D}}$ does not generally imply that subsequence combs $\mathbb{P}_C^{\mathsf{Y}_A \wr \mathsf{D}_A}$ and $\mathbb{P}_C^{\mathsf{Y}_B \wr \mathsf{D}_B}$ are equivalent with $|A| = |B|$.

4.5.3 Combs are the output of the "fix" operation

There is a relationship between combs and the "fix" operation defined in Richardson et al. (2017). In particular, suppose we have a probability $\mathbb{P}_\alpha$ and a comb $\mathbb{P}_\alpha^{\mathsf{Y}_{[2]}|\mathsf{D}_{[2]}}$. Then (assuming discrete sets)

$$\mathbb{P}_\alpha^{\mathsf{Y}_{[2]} \wr \mathsf{D}_{[2]}}(y_1, y_2|d_1, d_2) = \mathbb{P}_\alpha^{\mathsf{Y}_1|\mathsf{D}_1}(y_1|d_1)\mathbb{P}_\alpha^{\mathsf{Y}_2|\mathsf{D}_2}(y_2|d_2) \tag{4.190}$$

$$= \frac{\mathbb{P}_\alpha^{\mathsf{Y}_1|\mathsf{D}_1}(y_1|d_1)\mathbb{P}_\alpha^{\mathsf{D}_2|\mathsf{Y}_1\mathsf{D}_1}(d_2|y_1, d_1)\mathbb{P}_\alpha^{\mathsf{Y}_2|\mathsf{D}_2}(y_2|d_2)}{\mathbb{P}_\alpha^{\mathsf{D}_2|\mathsf{Y}_1\mathsf{D}_1}(d_2|y_1, d_1)} \tag{4.191}$$

$$= \frac{\mathbb{P}_\alpha^{\mathsf{Y}_{[2]}\mathsf{D}_2|\mathsf{D}_1}(y_1, y_2, d_2|d_1)}{\mathbb{P}_\alpha^{\mathsf{D}_2|\mathsf{Y}_1\mathsf{D}_1}(d_2|y_1, d_1)} \tag{4.192}$$

That is (at least in this case), the result of "division by a conditional probability" used in the fix operation is a comb. We speculate that the output of the fix operation is, in general, an n-comb, but we have not proven this.

4.6 Assessing decision problems for exchange commutativity

Exchange commutativity is a condition that, if it holds, allows a decision maker to use the map $D \rightarrow Y$ calculated from relative frequencies to determine the optimal course of action. The question is: when should a decision maker consider this assumption reasonable?, confronted with a decision problem actually adopt a causally contractible model $\mathbb{P}_C$ to help them make their decision? This is not an easy question for several reasons. Two of these are:

- The kind of symmetry required by exchange commutativity seems to us much harder to intuit than the kind of symmetry required by regular exchangeability

- The conditions of exchange commutativity and locality must hold for each choice in C

"Ordinary" exchangeability is often considered to be appropriate when modelling a measurement procedure that consists of a sequence of indistinguishable sub-procedures. A common example is a sequence of coin flips – there is (usually) no reason to consider any coin flip to differ in any important way from any other. Thus, one can reason, swapping the labels of the coin flips yields a measurement procedure that is effectively identical. It follows that the model should be unchanged under a permutation of the variables representing the sequence of flips – that is, it should be exchangeable[1]. The basic judgement call is then: the subprocedures for each coin flip are effectively identical.

Exchange commutativity requires a different kind of judgement. A common causal inference example features a decision procedure that yields a sequence of (treatment, outcome). Exchange commutativity asks us to compare the original procedure with an arbitrary procedure that shuffles the pairs. Then, *given any fixed vector of treatment values*, the resulting pair of procedures must be effectively indistinguishable. Full causal contractibility adds the requirement that, comparing two procedures of this type and restricting our attention to a subsequence of outcomes, we can ignore any differences between treatment vectors that do not correspond to the subsequence of interest.

This is not particularly easy to think about! Greenland and Robins (1986) mention the condition that the treatments of different patients could be swapped without changing the distribution over outcomes. This can be interpreted as saying: given two choices that induce deterministic treatment vectors, if the vector induced by the first is a permutation of the vector induced by the second, the resulting distributions of outcomes (appropriately permuted) should be identical. This is a consequence of exchange commutativity, but it is not equivalent: treatments (or "inputs") may not be deterministic for all choices, in which case it's not clear what "swapping treatments" means. If it's a hypothetical action that swaps treatments (see the discussion at the end of 3.1), it seems that some theory is needed to say what equivalence under such hypothetical actions imply for the actual choices to be evaluated.

A further complication is due to the fact that, by necessity, a probability set $\mathbb{P}_C$ models a measurement procedure for each of a set of choices C. Someone constructing a model $\mathbb{P}_C$ to help them deal with decision problem may want to reason that their state of knowledge after selecting some choice $\alpha \in C$ is the same as their state of knowledge when they are constructing $\mathbb{P}_C$. That is, they don't want to worry about whether their choice "depends on anything". The fact that they don't want to worry about this doesn't mean that they don't have to! The theory of probability sets is formal, and it can be augmented with decision rules to yield a formal theory of making decisions, but the correspondence between $\mathbb{P}_C$ and the "real things that constitute the decision problem"

[1] As Walley (1991, pg. 461) points out, the conclusion of exchangeability also requires the assumption that the measurement procedure should be modeled with a single probability distribution, which is an assumption that is being made in this chapter

is a judgement call, and it is possible to make poor calls. Example 4.6.1 is an example illustrating this. There are ways to deal with actions that "depend on things", see for example Gallow (2020)'s discussion of "managing the news", but the question of constructing appropriate models seems hard enough without the extra complication.

Individual-level causal contractibility is an attempt to specify a method for model construction that involves judgements that are (mostly) easier to think about than regular causal contractibility and that may sometimes yield regular causal contractibility as a result (Theorem 5.1.9). Notably, the assumption of individual-level causal contractibility can, under certain conditions, imply that a model is causally contractible conditional on an "unobserved" variable, analogous to the familiar assumption of hidden confounding.

4.6.1 Causal contractibility can be undermined by a lack of control

It is much easier to construct a model $\mathbb{P}_C$ if the choice C doesn't "depend on" anything. However, this property is not guaranteed, as Example 4.6.1 shows – the model need to be specified in the right way for the right kind of problem.

Example 4.6.1 (Confounded choices)**.** We set this up in terms of an "analyst" and an "administrator" not because it's necessary for the example but because it can help make it easier to understand. The analyst's job is to construct a model $\mathbb{P}_C$, evaluate different options $\alpha \in C$ and offer advice regarding the choice. The administrator's job is to actually choose some $\alpha \in C$ satisfying the analyst's advice and to carry out the associated procedure.

This separation of concerns gives the administrator a degree of freedom in their choice: they can potentially choose α with access to information that the analyst lacks.

In particular, suppose $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$, $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$, $\mathsf{U} = (\mathsf{U}_i)_{i\in\mathbb{N}}$ and the set of choices $C = [0,1]^{\mathbb{N}}$ is a length $\mathbb{N}$ sequence of probability distributions in $\Delta(\{0,1\})$. The analyst, based on their knowledge of the experiment, constructs the model $\mathbb{P}_C$ where $\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{U}_i\mathsf{D}_i}(1|\cdot,\cdot)$ is given by:

	$\mathsf{D}_i = 0$	$\mathsf{D}_i = 1$
$\mathsf{U}_i = 0$	0	0
$\mathsf{U}_i = 1$	1	1

and

$$\mathbb{P}_C^{\mathsf{U}_i}(\{u\}) = 0.5 \tag{4.193}$$

and

$$\mathbb{P}_\alpha^{\mathsf{D}_i}(1) = \alpha_i \tag{4.194}$$

where $\alpha = (\alpha_i)_{i\in\mathbb{N}}$ and the triples $(\mathsf{D}_i, \mathsf{U}_i, \mathsf{Y}_i)$ are mutually independent given C. From the analyst's point of view, $\mathbb{P}_C^{\mathsf{Y}|\mathsf{UD}}$ is causally contractible, and U_i is identically uniform for all i.

The analyst recommends any α such that $\lim_{n\to\infty}\sum_i^n \frac{\alpha_i}{n} = 0.5$ (acknowledging that, in this contrived example, there's no obvious reason to do so). Suppose that the administrator operates by the following rule: *first* they observe the binary result of $\mathcal{U}_i$, then they choose α_i equal to whatever they saw with an ϵ sized step towards 0.5. That is, if they see $\mathcal{U}_i \bowtie 1$, they choose $\alpha_i = 1-\epsilon$.

Then, from the analyst's point of view, $\alpha_i = 1-\epsilon \implies \mathbb{P}_{1-\epsilon}^{\mathsf{Y}_i|\mathsf{D}_i}(1|d) = 1$ for all i and $\alpha_j = \epsilon \implies \mathbb{P}_{1-\epsilon}^{\mathsf{Y}_j|\mathsf{D}_j}(1|d) = 0$ for all j. This means that $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ is not exchange commutative. The administrator's rule for choosing α means that, though the analyst does not know the outcome of $\mathcal{U}$, they know what it would be for any α.

Example 4.6.1 features a situation where it is arguable whether or not the analyst was applying $\mathbb{P}_C$ to a "decision problem". In particular, they did not offer a particular choice α as a result of their analysis, but instead offered a subset of C. For example, the analyst might require that . Example 4.6.1 in fact satisfies this requirement.

The original justification for having a set of choices C is that C is the set of things that, after deliberation aided by the model $\mathbb{P}_C$, the decision maker might select. Example 4.6.1 does not satisfy this understanding of the meaning of the set C – the analyst aided by $\mathbb{P}_C$ selects nothing or, in the previous paragraph, selects a subset of C rather than an element of C. Thus this example suggests that one should be cautious about using a probability set $\mathbb{P}_C$ to evaluate choices without choosing anything definite based on this evaluation.

Kasy (2016) argues that "randomised controlled trials are not needed for causal identifiability, only controlled trials", and suggests that experiments should sometimes be designed with deterministic assignments of patients to treatment and control groups. From one point of view, if $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ is causally contractible, then causal contractibility holds whether the choice α yields a deterministic or nondeterministic distribution over D. If the same model $\mathbb{P}_C$ is used to design a 'trial' (say, for the first n $(\mathsf{D}_i, \mathsf{Y}_i)$ pairs) and to select optimal actions for the pairs after this, there is no obvious reason to choose the first n actions according to a random number generator – one may be able to get better coverage of the actions of interest by choosing deterministically. Indeed, as Kasy (2016) points out, if random choices are mixtures of deterministic choices, then the random choices can be at best as good as some deterministic choice (in his paper, Kasy considers situations with covariates, which we ignore for simplicity).

However, a typical trial involves an experimenter conducting the trial (which comprises the procedure for "the first n pairs") and publishing a paper summarising their findings. A reader may then read the paper and use the results to make decisions (which comprises the procedure for "the pairs after the first n"). It is the reader who is in the position of needing a model $\mathbb{P}_C$ accounting for both the experimental data and the data that arises as a consequence of her actions. In this case, the reader is in a position somewhat like the "analyst" in Example 4.6.1, while the experimenter is the "administrator" from this example. Fixing the proportion of treatments in each class (like in Example 4.6.1) leaves the experimenter with many degrees of freedom to select deterministic sequences of actions. On the other hand, if the experimenter may only select a target proportion of treatments, and must then assign treatments randomly given this target, they have much less freedom to influence the consequences of the experiment. An experimenter with many degrees of freedom might well be trusted to make choices in a manner that is transparent to any readers, but this does require an additional assumption on the part of the readers.

Chapter 5

Other causal modelling frameworks

5.1 Weaker assumptions than causal contractibility

The results so far apply to purely observational models or to models where every "input" in the sequence is fixed at the point of choosing α (or a fixed random function is chosen at this point). Most of the interest in causal inference is how to use observational data – which is plentiful – to deduce consequences of choices. Suppose in the following that superscript "o" refers to observational variables (obtained by some measurement procedure not responsive to choices) and "v" refers to interventional variables (obtained by some measurement procedure responsive to choices). That is $\mathsf{Y}^o := (\mathsf{Y}_i^o)_{i\in\mathbb{N}}$ is a sequence of observational variables, Y^v a sequence of interventional variables and $\mathsf{Y}^{o,v} := (\mathsf{Y}_i^o, \mathsf{Y}_i^v)_{i\in\mathbb{N}}$ is a mixed sequence of both observational and interventional variables. Y_i^o and Y_i^v are assumed to take values in the same set Y.

One approach to bridging the gap between observations and interventions is to assume "causal sufficiency", which is tantamount (in the data-independent case) to assuming causal contractibility of $\mathbb{P}_C^{\mathsf{Y}^{o,v}|\mathsf{X}^{o,v}\mathsf{D}^{o,v}}$ with D^v responsive to choices and X^v unresponsive (see Example 4.3.17). As discussed, this is rarely a reasonable assumption – it implies interchangeability between observational and interventional samples.

A weaker assumption that is often adopted is to consider models satisfying causal contractibility with respect to $\mathbb{P}_C^{\mathsf{Y}^{o,v}|\mathsf{U}^{o,v}\mathsf{D}^{o,v}}$, where $\mathsf{U}^{o,v}$ is unobserved. That is, while $\mathsf{U}^{o,v}$ appears in the model, it is not associated with any measurement procedure. This model still asserts that $(\mathsf{U}_i^o, \mathsf{X}_i^o, \mathsf{Y}_i^o)$ triples are interchangeable with $(\mathsf{U}_i^v, \mathsf{X}_i^v, \mathsf{Y}_i^v)$ triples, but neither of these are measurement outcomes. On the other hand, $(\mathsf{D}_i^o, \mathsf{Y}_i^o)$ pairs are not generally interchangeable with $(\mathsf{D}_i^v, \mathsf{Y}_i^v)$.

Consider models that satisfy causal contractibility with respect to $\mathbb{P}_C^{\mathsf{Y}^{o,v}|\mathsf{W}^{o,v}}$, where no comment is made about whether $\mathsf{W}^{o,v}$ is observed, unobserved or some function of observed and unobserved variables. This is a generalisation of the class of models discussed in the previous paragraph. In isolation, this assumption is not especially interesting – for example, the support of W_i^o and W_i^v might be disjoint. Suppose also, then, that W is finite and W_i^o has full support. This assumption amounts to the assumption that, no matter what choice is made, "nothing truly new can be done" (which we call "Ecclesiastes' assumption"[1]). More precisely, for any choice $\alpha \in C$ and any consequence Y_i^v, there is a random subsequence Q of indices $(1, 2, 3, ...)$ such that the distribution $\mathbb{P}_\alpha^{\mathsf{Y}^{o,v}}$ is unchanged by permutations that only swap elements in the sequence $(RVY_{\mathsf{Q}}^o, \mathsf{Y}_i^v)$.

Theorem 5.1.1. *Given just-do model $\mathbb{P}_C$ with $\mathbb{P}_C^{\mathsf{Y}^{o,v}|\mathsf{W}^{o,v}}$ causally contractible, W finite and $\mathbb{P}_C^{\mathsf{W}^o|\mathsf{H}}(w|h) > 0$ for all w, h, define $q : W^{\mathbb{N}} \times W \to (\{*\} \cup \mathcal{P}(\mathbb{N})$ by*

$$q : ((w_j^o)_{\mathbb{N}}, w_i^v) \mapsto \{j | w_j^o = w_i^v\} \tag{5.1}$$

and take $\mathsf{Q} := q \circ (\mathsf{W}^o, \mathsf{W}_i^v)$ for arbitrary $i \in \mathbb{N}$. For an index set $U \in \mathbb{N}$ take $swap_. : Y^{\mathbb{N}} \times Y^{\mathbb{N}} \to Y^{\mathbb{N}} \times Y^{\mathbb{N}}$ to be an arbitrary finite swap that acts as the identity on all indices $(j, x) \notin \mathsf{Q} \times \{o\} \cup \{(i, v)\}$. Then $\mathbb{P}^{\mathsf{Y}^o\mathsf{Y}_i^v} swap_{\mathsf{Q}} = \mathbb{P}^{\mathsf{Y}^o\mathsf{Y}_i^v}$.

Proof. Note that for $B_j \in \mathcal{W}$, where $\rho_q : \mathbb{N} \times \{i, v\} \to \mathbb{N} \times \{i, v\}$ is the permutation function associated with swap_q

$$\mathbb{P}_\alpha^{\mathsf{W}^o\mathsf{W}_i^v} \mathrm{swap}_{\mathsf{Q}}(\bigtimes_{j \in \mathbb{N}} B_j) = \int_{W^{\mathbb{N}}} \int_{\mathcal{P}(\mathbb{N})} \prod_{k \notin q \times \{o\} \cup \{(i,v)\}} \delta_{w_k}(B_k) \prod_{l \in q \times \{o\} \cup \{(i,v)\}} \delta_{\rho_q(w_l)}(B_l) \mathbb{P}_\alpha^{\mathsf{Q}|\mathsf{W}^o\mathsf{W}_i^v}(\mathrm{d}q|w) \mathbb{P}_\alpha(\mathrm{d}w) \tag{5.2}$$

$$= \int_{W^{\mathbb{N}}} \int_{\mathcal{P}(\mathbb{N})} \prod_{k \notin q \times \{o\} \cup \{(i,v)\}} \delta_{w_k}(B_k) \prod_{l \in q \times \{o\} \cup \{(i,v)\}} \delta_{w_l}(B_l) \mathbb{P}_\alpha^{\mathsf{Q}|\mathsf{W}^o\mathsf{W}_i^v}(\mathrm{d}q|w) \mathbb{P}_\alpha(\mathrm{d}w) \tag{5.3}$$

$$= \mathbb{P}_\alpha^{\mathsf{W}^o\mathsf{W}_i^v} \tag{5.4}$$

where Eq. 5.3 follows from the fact that for every $k, l \in q \times \{o\} \cup \{(i, v)\}$, $w_k = w_l$.

Thus for $A \in \mathcal{Y}^{\mathbb{N}}$

$$\mathbb{P}_\alpha^{\mathsf{Y}^o\mathsf{Y}_i^v} \mathrm{swap}_{\mathsf{Q}}(A) = [\mathbb{P}_\alpha^{\mathsf{W}^o\mathsf{W}_i^v} \mathbb{P}_\alpha^{\mathsf{Y}^o\mathsf{Y}_i^v|\mathsf{Q}\mathsf{W}^o\mathsf{W}_i^v} \mathrm{swap}_{\mathsf{Q}}](A) \tag{5.5}$$

$$= [\mathbb{P}_\alpha^{\mathsf{W}^o\mathsf{W}_i^v} \mathrm{swap}_{\mathsf{Q}^{-1}} \mathbb{P}_\alpha^{\mathsf{Y}^o\mathsf{Y}_i^v|\mathsf{W}^o\mathsf{W}_i^v} \mathrm{swap}_{\mathsf{Q}}](A) \tag{5.6}$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}^o\mathsf{Y}_i^v} \tag{5.7}$$

Where Eq. 5.7 follows from causal contractibility of $\mathbb{P}_\alpha^{\mathsf{Y}^o\mathsf{Y}_i^v|\mathsf{W}^o\mathsf{W}_i^v}$. □

[1]Ecclesiastes 1:9 reads "Everything that happens has happened before; nothing is new, nothing under the sun."(noa, 1995)

It also follows from Ecclesiastes' assumption and finite W that if some X_i^o, Z_i^o are *deterministically* related given W, then $\mathbb{P}_C^{\mathsf{Z}|\mathsf{X}}$ is causally contractible.

Theorem 5.1.2. *Given just-do model* $\mathbb{P}_C$ *with* $\mathbb{P}_C^{\mathsf{X}^{o,v}\mathsf{Z}^{o,v}|\mathsf{W}^{o,v}}$ *causally contractible,* W *finite and* $\mathbb{P}_C^{\mathsf{W}_i^o|\mathsf{H}}(w|h) > 0$ *for all* w, h*, if* $\mathbb{P}_C^{\mathsf{Z}_0^o|\mathsf{X}_0^o\mathsf{H}}$ *is deterministic then* $\mathbb{P}_C^{\mathsf{Z}^{o,v}|\mathsf{X}^{o,v}}$ *is causally contractible.*

Proof. Because $\mathbb{P}_\alpha^{\mathsf{W}_0^o}\mathbb{P}_C^{\mathsf{Z}_0^o|\mathsf{X}_0^o\mathsf{W}_0^o\mathsf{H}}$ is deterministic, so is $\mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{W}_0\mathsf{H}}$.

Fix $h \in H$. Suppose there is some $w, w' \in W$ such that

$$\mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{W}_0\mathsf{H}}(A|x, w, h) \neq \mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{W}_0\mathsf{H}}(A|x, w', h) \tag{5.8}$$

then, by determinism, we can assume without loss of generality

$$\mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{W}_0\mathsf{H}}(A|x, w, h) = 1 \tag{5.9}$$

$$\mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{W}_0\mathsf{H}}(A|x, w', h) = 0 \tag{5.10}$$

but W is finite and $\mathbb{P}_C^{\mathsf{W}_i^o|\mathsf{H}}(w|h) > 0$ for all w, so there is some $a > 0$ such that $\mathbb{P}_C^{\mathsf{W}_i^o|\mathsf{H}}(w|h) \geq a$ for all w, and so

$$a \leq \sum_{w \in W} \mathbb{P}_C^{\mathsf{W}_0^o|\mathsf{X}_0^o,\mathsf{H}}(w|x, h)\mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{W}_0\mathsf{H}}(A|x, w, h) \leq 1 - a \tag{5.11}$$

contradicting determinism of $\mathbb{P}_C^{\mathsf{Z}_0^o|\mathsf{X}_0^o\mathsf{H}}$.

Thus for all w, w'

$$\mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{W}_0\mathsf{H}}(A|x, w, h) = \mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{W}_0\mathsf{H}}(A|x, w', h) \tag{5.12}$$

i.e. $\mathsf{Z}_0 \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{W}_0, \mathsf{C})|(\mathsf{X}_0, \mathsf{H})$. But then there is some $\mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{H}}$ such that

$$\mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{W}_0\mathsf{H}} = \mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{H}} \otimes \text{erase}_W \tag{5.13}$$

$$\implies \mathbb{P}_\alpha^{\mathsf{Z}_i^v|\mathsf{X}_i^v\mathsf{H}} = \mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{H}} \tag{5.14}$$

□

Theorem 5.1.2 doesn't hold in the case of approximate determinism, however. Intuitively, approximate determinism can hold if there is some value of W for which Z is not conditionally independent given H and X, but it only ocurrs very rarely in observations. On the other hand, values of W rare in observations might, under some choices, become common.

Example 5.1.3. Say $\mathbb{P}_C^{\mathsf{Z}_i^o|\mathsf{X}_i^o\mathsf{H}}$ is *approximately deterministic* if $\mathbb{P}_C^{\mathsf{Z}_i^o|\mathsf{X}_i^o\mathsf{H}}(A|x, h) \in [0, \epsilon] \cup [1 - \epsilon, 1]$ for all $A \in \mathcal{Z}$, $x, h \in X \times H$.

Take $Z = X = W = H = \{0, 1\}$. Set

$$\mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{W}_0\mathsf{H}}(1|1, 1, 1) = 1 \tag{5.15}$$

$$\mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{W}_0\mathsf{H}}(1|1, 0, 1) = 0 \tag{5.16}$$

and

$$\mathbb{P}_C^{\mathsf{W}_0^o|\mathsf{H}}(1|1) = 1 - \epsilon \tag{5.17}$$

then

$$\mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{H}}(1|1,1) = 1 - \epsilon \tag{5.18}$$

however, suppose there is some α such that

$$\mathbb{P}_\alpha^{\mathsf{W}_i^v|\mathsf{H}}(1|1) = 0 \tag{5.19}$$

then

$$\mathbb{P}_\alpha^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{H}}(1|1,1) = 0 \tag{5.20}$$

$$\neq \mathbb{P}_C^{\mathsf{Z}_0|\mathsf{X}_0\mathsf{H}}(1|1,1) \tag{5.21}$$

5.1.1 Individual-level causal contractibility

Individual-level causal contractibility is our attempt to formulate a set of conditions sufficient for causal contractibility that may be easier to think about than the bare assumption. It is inspired by the fact that many treatments of causal identifiability refer to sequences of "individuals", "patients", "units" or the like, all of which are considered to be "essentially identical" from the point of view of the decision maker. However, these individuals are not directly referenced by the models under consideration. What we do here is consider models that *do* reference individuals – in particular, we suppose that each sub-procedure produces an input variable, an output variable and a *unique identifier* (which can be thought of as something like someone's license number and state of issue). Under some conditions, assuming causal contractibility with respect to inputs, identifiers and outputs can imply causal contractibility with respect to inputs and outputs only. This clearly hasn't solved the entire problem of assessing a problem for causal contractibility, as a causal contractibility assumption is required to get off the ground.

it might still be helpful, though? I don't really know what to say

References to individual-level causal contractibility

The role of individuals has often been mentioned in literature on causal inference. 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.

Here, the idea of "exchangeable individuals" plays a role in the author's reasoning about model construction, but "individuals" are not actually referenced by the resulting model, and "exchanging individuals" does not correspond to a model transformation.

Dawid (2020) suggests (with some 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.

As in the previous work, the similarity of individuals involved in an experiment is raised when justifying particular model constructions, but the individuals are not referenced by the model.

Pearl (2009, pg. 98) writes

> Although the term unit in the potential-outcome literature normally stands for the identity of a specific individual in a population, a unit may also be thought of as the set of attributes that characterize that individual, the experimental conditions under study, the time of day, and so on all of which are represented as components of the vector u in structural modeling.

Once again, the idea of an individual (or a particular set of conditions) is raised in the context of explaining modelling choices. Unlike the previous authors, Pearl introduces a vector u to stand for the "unit". However, he subsequently assumes that u is a sequence of *independent samples* from some distribution. This seems to contradict an important feature of "individuals" or "units": individuals are typically supposed to be unique, a property that will usually not be satisfied by independently sampling from some distribution (at least, as long as the distribution is discrete).

Finally, Rubin (2005) writes:

> Here there are N units, which are physical objects at particular points in time (e.g., plots of land, individual people, one person at repeated points in time).

Note that Rubin's explanation of *units* guarantees that they are unique: they are particular things at particular times. These units are associated with input-output functions (the *potential outcomes*), which are later assumed to be exchangeable:

> the indexing of the units is, by definition, a random permutation of $1, ..., N$, and thus any distribution on the science must be row-exchangeable

Our proposition is: can the intuition that unique individuals are an important for the motivation for causal models, be captured by considering models that feature "unique identifier" variables referencing these unique individuals?

Unique identifiers

A sequence of *unique identifiers* is a vector of finite or infinite length such that no two coordinates are equal. We are interested in models that assign positive probability to any particular coordinate having any particular value. This is straightforward in the finite case. In the infinite case, note that a vector of $|\mathbb{N}|$ unique values with an arbitrary entry k in the jth coordinate can be obtained by starting with $(i)_{i\in\mathbb{N}}$ and then transposing j with k. More generally, we consider infinite length sequences of unique identifiers to be elements of the set of finite permutations $\mathbb{N}\to\mathbb{N}$.

Definition 5.1.4 (Measurable space of unique identifiers)**.** The measurable space of unique identifiers $(I,\mathcal{I})$ is the set I of finite permutations $\mathbb{N}\to\mathbb{N}$ with the discrete σ-algebra $\mathcal{I}$.

The set I is countable, as it is the countable union of finite subsets (i.e. the permutations that leave all but the first n numbers unchanged for all n).

Definition 5.1.5 (Unique identifier)**.** Given a sample space $(\Omega,\mathcal{F})$, a *sequence of unique identifiers* $\mathcal{I}:\Omega\to I$ is a variable taking values in I.

The values of each coordinate of sequence of unique identifiers is just called an identifier (for obvious reasons, we don't call it an identity).

Definition 5.1.6 (Identification)**.** Given I, define the i-th *identifier* $\mathsf{I}_i = \mathrm{ev}(i,\mathsf{I})$, where $\mathrm{ev}:\mathbb{N}\times I\to\mathbb{N}$ is the evaluation map $(i,f)\mapsto f(i)$.

For *any* sample space $(\Omega,\mathcal{F})$ we can define a trivial $\mathcal{I}$ that maps every $\omega\in\Omega$ to $(1,2,3,....) =: (\mathbb{N})$. In this case, the identifiers are all known by the modeller at the outset. Using this sequence of identifiers renders exchange commutativity trivial, and isn't of much interest to us.

Example 5.1.7. Given a sequential just-do model $(\mathbb{P}_C,(\mathsf{D},\mathsf{I}),\mathsf{Y})$ where I is the identifier variable $\omega\mapsto(\mathbb{N})$, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{DI}}$ commutes with exchange.

This is because for any permutation $\rho:\mathbb{N}\to\mathbb{N}$ except the identity, $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{DI}}$ and $\mathrm{swap}_\rho\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{DI}}$ will have no common support; the first will be supported on $\mathsf{I}\bowtie(\mathbb{N})$ only, and the second only on $\mathsf{I}\bowtie\rho(\mathbb{N})$.

Individual-level causal contractibility and unobserved confounding

Our first result is that some models with individual-level causal contractibility can be seen as models with unobserved confounding. A model $\mathbb{P}_C$ with individual-level causal contractibility features a causally contractible $\mathbb{P}_C^{\mathsf{Y}|\mathsf{ID}}$ for a sequence of outputs Y, inputs D and individual identifiers I. A model with unobserved confounding features causally contractible $\mathbb{P}_C^{\mathsf{Y}|\mathsf{UD}}$ where Y and D are as before and U is an "unobserved confounder". They key difference between I and U is that the individual identifier for each observation is unique, while unobserved variables (typically) have $|U|<N$ where N is the number of observations.

Individual-level causal contractibility and ordinary causal contractibility

Our second key result is that individual-level causal contractibility along with the assumptions of exchangeability of individuals and sufficient control of inputs implies causal contractibility with respect to inputs and outputs.

So, a judgement of symmetry among sub-experiments is not enough for causal contractibility. What is enough?

In the following, it is helpful to assume that each sub-experiment has a "unique identifier" I_i, with the sequence of all sub-experiment labels given by I. With this, if $\mathbb{P}_C^{\mathsf{Y}|\mathsf{DI}}$ is assumed causally contractible, then it's possible to talk about the individual response functions $\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{I}_i\mathsf{HD}_i}$. These plays a role very similar to the ith vector of potential outcomes Y_i^D. Because I_i is unique (i.e. never equal to I_j for $j \neq i$), only one observation of any individual is ever given, just like only one potential outcome is ever observed.

Theorem 5.1.9 presents a set of sufficient conditions for causal contractibility of $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$:

1. There exist variables I representing "unique experiment identifiers" which satisfy the assumption that $\mathbb{P}_C^{\mathsf{Y}|\mathsf{DI}}$ is causally contractible (informally: it doesn't matter which order the experiments are conducted in, and treatments in each experiment do not affect any other experiments)

2. The identifiers themselves are not informative regarding outcomes: $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{I}|\mathsf{C})$

3. The inputs D are substitutable for the choice C with respect to Y and I: $\mathsf{YI} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{C}|\mathsf{D}$ and $\mathsf{YI} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{D}|\mathsf{C}$

However, as we will show, even these conditions can be subtle to assess.

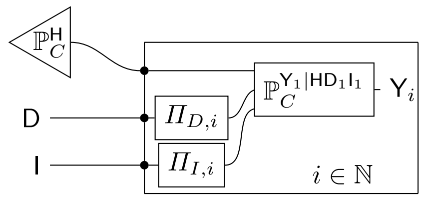
As an example of the application of Theorem 5.1.9, consider an experiment where n patients, each with an individual identifier I_i, receive treatment D_i and experience outcome Y_i. $\mathbb{P}_C^{\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{I}_{[n]}}$ can be extended to an infinite sequence $\mathbb{P}_C^{\mathsf{Y}|\mathsf{DI}}$ that is causally contractible (see Assumption 1), and no matter which choice $\alpha \in C$ is given, all identifiers can be swapped without altering the distribution over consequences (see Assumption 2), and finally that the treatment vector D is a deterministic and invertible function of the choice $\alpha \in C$ then $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ is causally contractible, and hence there are response functions $\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}$.

Theorem 5.1.9 can also be extended to the case where D is a function of the choice α and a "random signal" R.

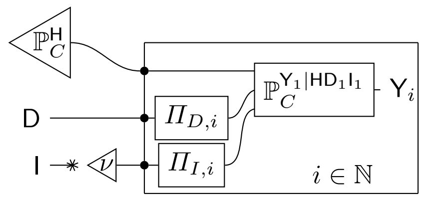
Lemma 5.1.8. *Given sequential just-do model* $(\mathbb{P}_C, (\mathsf{D}, \mathsf{I}), \mathsf{Y})$ *with* $\mathbb{P}_C^{\mathsf{Y}|\mathsf{DI}}$ *causally contractible, if* $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{I}, \mathsf{C})|\mathsf{D}$ *and for any* $j \in I$, $\sum_{\alpha\in C} \mathbb{P}_\alpha^{\mathsf{I}_i}(j) > 0$, *then* $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ *is also causally contractible.*

Proof. For arbitrary $\nu \in \Delta(I^{\mathbb{N}})$ such that $\sum_{\alpha \in C} \mathbb{P}_\alpha^{\mathsf{I}_i} \gg \nu$, by assumption of causal contractibility of $\mathbb{P}_C^{\mathsf{Y}|\mathsf{DI}}$ and Theorem 4.3.11

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{DI}} \overset{\mathbb{P}_C}{\cong} \tag{5.22}$$



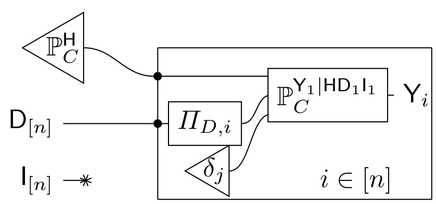
$$\overset{\mathbb{P}_C}{\cong} \tag{5.23}$$



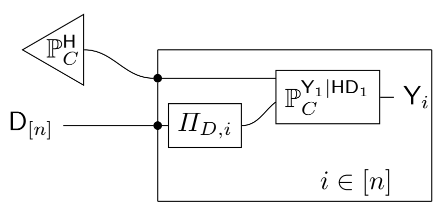
Where $\Pi_{D,i} : D^{\mathbb{N}} \to D$ projects the ith coordinate, and similarly for $\Pi_{Y,i}$.

In particular, for any $i \in \mathbb{N}$, $j \in I$, this holds for some ν such that $\nu(\Pi_{Y,i}^{-1}(j) = 1$ and by extension for any finite $A \subset \mathbb{N}$ we can find ν such that $\nu(\Pi_{Y,i}^{-1}(j) = 1$ for all $i \in A$, $j \in I$. Thus for any $n \in \mathbb{N}$

$$\mathbb{P}_C^{\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{I}_{[n]}} \overset{\mathbb{P}_C}{\cong} \tag{5.24}$$



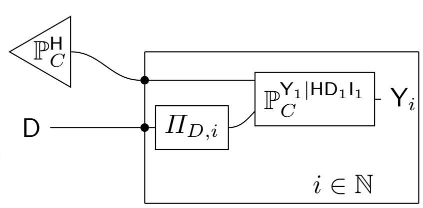
$$\overset{\mathbb{P}_C}{\cong} \tag{5.25}$$



where Equation 5.24 follows from Theorem 2.3.6 and Equation 5.25 follows from the fact that Equation 5.24 holds for arbitrary $j \in I$.

Thus by Lemma 4.3.5

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}} = \tag{5.26}$$



Applying Theorem 4.3.11, $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ is causally contractible. $\square$

Theorem 5.1.9. *Given a sequential just-do model* $(\mathbb{P}_C, (\mathsf{D}, \mathsf{I}), \mathsf{Y})$ *on* $(\Omega, \mathcal{F})$ *with* Y *standard measurable and* C *countable,* $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{DI}}$ *causally contractible for each* α*, if*

$$\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{I}|\mathsf{C} \tag{5.27}$$

$$\mathsf{YI} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{C}|\mathsf{D} \tag{5.28}$$

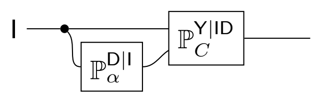
$$\mathsf{YI} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{D}|\mathsf{C} \tag{5.29}$$

$$\forall i, j \in \mathbb{N} : \sum_{\alpha \in C} \mathbb{P}_\alpha^{\mathsf{I}_i}(j) > 0 \tag{5.30}$$

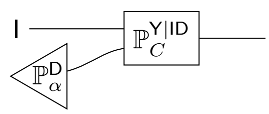
then $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ *is causally contractible.*

Proof. For any $\alpha \in C$

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{I}} = \tag{5.31}$$

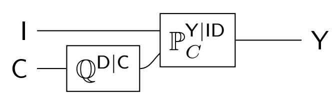


$$= \tag{5.32}$$



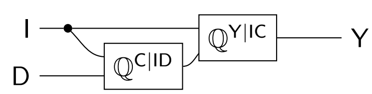
Define $\mathbb{Q}$ by $\alpha \mapsto \mathbb{P}_\alpha$ and $\mathbb{Q}^{\cdot|\cdot\mathsf{C}}$ by $\alpha \mapsto \mathbb{P}_\alpha^*$ and $\mathbb{Q}^{\mathsf{C}}$ is an arbitrary distribution in $\Delta(C)$ with full support. Note that the support of $\mathbb{Q}^{\mathsf{IDY}}$ is the union of the support of $\mathbb{P}_\alpha^{\mathsf{IDY}}$ for all α. Then

$$\mathbb{Q}^{\mathsf{Y}|\mathsf{IC}} \overset{\mathbb{Q}}{\cong} \tag{5.33}$$

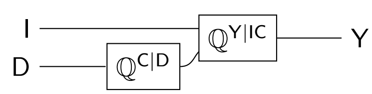


By assumption $\mathsf{YI} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{D}|\mathsf{C}$, it is also the case that

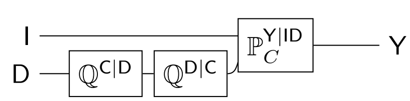
$$\mathbb{Q}^{\mathsf{Y}|\mathsf{ID}} \overset{\mathbb{Q}}{\cong} \tag{5.34}$$



$$\overset{\mathbb{Q}}{\cong} \tag{5.35}$$



$$\overset{\mathbb{Q}}{\cong} \tag{5.36}$$

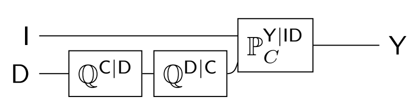


But

$$\mathbb{Q}^{\mathsf{Y}|\mathsf{ID}} = \sum_{\alpha \in C} \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{ID}} \mathbb{Q}^{\mathsf{C}}(\alpha) \tag{5.37}$$

$$= \mathbb{P}_C^{\mathsf{Y}|\mathsf{ID}} \tag{5.38}$$

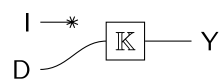
$$\implies$$



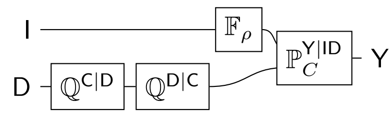
$$= \mathbb{P}_C^{\mathsf{Y}|\mathsf{ID}} \tag{5.39}$$

Furthermore, by assumption $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{I}|\mathsf{C}$, so there is some $\mathbb{K} : C \rightarrow Y$ such that

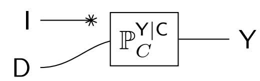
$$\mathbb{Q}^{\mathsf{Y}|\mathsf{IC}} \overset{\mathbb{Q}}{\cong} \tag{5.40}$$



$$\implies \mathbb{P}_C^{\mathsf{Y}|\mathsf{ID}} = \tag{5.41}$$



$$= \tag{5.42}$$



Then by Lemma 5.1.8, $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ is causally contractible. □

Theorem 5.1.9 can be extended to the case where decisions D are a one-to-one deterministic function of the choice, or a random mixtures of one-to-one deterministic functions of the choice.

Theorem 5.1.10. *Consider a sequential just-do model* $(\mathbb{P}_{C'}, \mathsf{D}, \mathsf{Y})$ *where* $\mathbb{P}_{C'}^{\mathsf{Y}|\mathsf{D}}$ *is causally contractible, and construct a second model* $(\mathbb{P}_C, \mathsf{D}, \mathsf{Y})$ *where* $\mathbb{P}_C$ *is the union of* $\mathbb{P}_{C'}$ *and its convex hull. Then* $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ *is also causally contractible.*

Proof. For all $\alpha \in C$, there is some probability measure $\mu : C' \to [0,1]$ such that

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}} = \sum_{\beta \in C'} \mu(\beta) \mathbb{P}_\beta^{\mathsf{Y}|\mathsf{D}} \tag{5.43}$$

$$= \mathbb{P}_{C'}^{\mathsf{Y}|\mathsf{D}} \tag{5.44}$$

thus

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}} = \mathbb{P}_{C'}^{\mathsf{Y}|\mathsf{D}} \tag{5.45}$$

and in particular, $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ is causally contractible. □

The assumption $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{I}|\mathsf{C}$ can be understood in terms of *permutability of identifiers*. An *identifier variable* is a variable I that takes values in the set of finite permutations of $\mathbb{N}$. It is associated with a sequence $(\mathsf{I}_i)_{i \in \mathbb{N}}$ where $\mathsf{I}_i = \mathsf{I}(i)$. Each I_i takes values in $\mathbb{N}$ and $\mathsf{I}_i \neq \mathsf{I}_j$ for all $j \neq i$.

Definition 5.1.11 (Identifier variable)**.** Given a probability set $\mathbb{P}_C$ on $(\Omega, \mathcal{F})$, let I be the set of finite permutations $\mathbb{N} \to \mathbb{N}$. A variable $\mathsf{I} : \Omega \to I$ be a variable taking values in I is an *identifier variable*.

If a uniform conditional probability is invariant to permutations of an index variable, then it is independent of that index variable.

Lemma 5.1.12. *Given a probability set $\mathbb{P}_C$ where $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{C}|(\mathsf{D},\mathsf{I})$ and $\mathsf{I}: \Omega \to I$ is an identifier variable, if for each finite permutation $\rho : \mathbb{N} \to \mathbb{N}$*

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{I}} = (swap_{\rho(I)} \otimes Id_X)\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{I}} \tag{5.46}$$

then $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{I}|\mathsf{C}$.

Proof. By definition of the set I of finite permutations, for every $\rho \in I$, $B \in \mathcal{Y}^{\mathbb{N}}$, $d \in D^{\mathbb{N}}$ there is a finite permutation $\rho^{-1} \in I$ such that $\rho \circ \rho^{-1} = \mathrm{id}_{\mathbb{N}}$. Then

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{I}}(B|\rho) = (\mathbb{F}_{\rho^{-1}} \otimes \mathrm{Id}_X)\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{I}}(B|\rho) \tag{5.47}$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{I}}(B|\mathrm{id}_{\mathbb{N}}) \tag{5.48}$$

Therefore

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{I}} \overset{\mathbb{P}_C}{\cong} \mathrm{erase}_I \otimes \mathbb{P}_\alpha^{\mathsf{Y}} \tag{5.49}$$

□

Theorem 5.1.10 can be used to argue that, given a sequence of experiments causally contractible under deterministic choices, adding random mixtures of these choices also yields a causally contractible sequence. Kasy (2016) argues that as long as the experimenter controls the treatment assignment, causal effects are identified (i.e. the randomisation step is not strictly necessary). Example 5.1.13 shows that this argument might be supported, but Example 5.1.14 shows that there are subtle ways that might lead to this argument failing. We consider a simpler case than Kasy (2016), where there are no covariates to worry about.

We assume an infinite sequence, which is clearly unreasonable. Extending the representation theorems to the case of finite sequences, using for example the result of Diaconis and Freedman (1980) with establishes that finite exchangeable distributions are approximately mixtures of independent and identically distributed sequences, would allow some implausible assumptions in the following example to be removed.

Example 5.1.13. A sequential experiment is modeled by a probability set $\mathbb{P}_C$ with binary treatments $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and binary outcomes $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$. The set of choices C is the set of all probability distributions$\Delta(D^N)$ for some $N \subset \mathbb{N}$ (this is to ensure C is countable).

Each treatment D_i is given to a patient, and each patient provides a unique identifier I_i which for simplicity we assume is a number in $\mathbb{N}$ (instead of, say, a driver's license number and state of issue), and that (implausibly) there is a positive probability for I_i to take any value in $\mathbb{N}$ for any choice α.

The treatments are decided as follows: the analyst consults the model $\mathbb{P}_C$, and, according to $\mathbb{P}_C$ and some previously agreed upon decision rule, comes up with a possibly stochastic sequence of treatment distributions $\alpha := (\mu_i)_{i\in N}$ with each μ_i in $\Delta(\{0,1\})$. If μ_i is deterministic – that is, it puts probability

1 on some treatment d_i, the experiment administrator will assign patient i the treatment d_i. Otherwise, if μ_i is nondeterministic, the administrator will consult some agreed-upon random number generator that yields treatment assignments according to μ_i.

Let $C' \subset C$ be the deterministic elements of C, and assume that all elements of C are a convex combination of elements of C'. The randomisation procedure is deemed sufficient to ensure that for any mixed $\alpha \in C$ where $\alpha = \sum_{\beta\in C'} \mu(\beta)\beta$, $\mathbb{P}_\alpha = \sum_{\beta\in C'} \mathbb{P}_\beta$.

Furthermore, assume $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{DI}}$ is causally contractible for each α. As discussed in Theorem 4.3.7, this means that for any $A, B \subset \mathbb{N}$, $|A| = |B|$, $\mathbb{P}_\alpha^{\mathsf{Y}_A|\mathsf{D}_A\mathsf{I}_A} = \mathbb{P}_C^{\mathsf{Y}_B|\mathsf{D}_B\mathsf{I}_B}$ and $\mathsf{Y}_A \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{D}_{A^C}, \mathsf{I}_{A^C})|(\mathsf{D}_A, \mathsf{I}_A, \mathsf{C})$. Roughly: holding any subsequence of treatments and individuals fixed, the joint distribution of consequences is the same no matter where they appear in the sequence of experiments and no matter what treatments or individuals appear elsewhere.

The analyst constructing the model has no particular knowledge about any identifier, and so *for any choice* reasons the associated model should be invariant to permutations of identifiers. This implies $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{I}|\mathsf{C}$ (see Lemma 5.1.12). The assumption that this holds given any choice can be tricky – not only must the identifiers appear symmetric to the analyst constructing the model, but nothing breaking this symmetry may be learned from the choice α, see the Example 5.1.14 next and the discussion afterwards. In this case, the reasoning is supported by the fact that the rule for selecting α is known in advance.

For the deterministic subset $C' \subset C$, $\mathsf{YI} \perp\!\!\!\perp^e_{\mathbb{P}_{C'}} \mathsf{D}|\mathsf{C}$ as D is deterministic for all elements of C', and $\mathsf{YI} \perp\!\!\!\perp^e_{\mathbb{P}_{C'}} \mathsf{C}|\mathsf{D}$ is also because restricted to this subset, D is a one-to-one function of C. By application of Theorem 5.1.9, $\mathbb{P}_{C'}^{\mathsf{Y}|\mathsf{D}}$ is causally contractible, and by application of Corollary 5.1.10, so is $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$

Permutability of identifiers can fail when the rule for selecting α is not known in advance. The following example is extreme in order to illustrate the issue clearly. The distinction between the analyst and the administrator is also intended to make the example easier to parse. The key point is that, when the rule for selecting α is not known in advance, symmetries that are apparent at the time of model construction do not necessarily hold for every choice α, and this remains true if e.g. the selection of choices leads to less extreme confounding or the analyst and the administrator are actually the same person.

The following example involves the choice α depending on some covariate U. It is not straightforward to express the idea that "α depends on U" in a probability set model $\mathbb{P}_C$, and they are intended to apply to situations where the choice doesn't depend on anything not already expressed in the model (as in Example 5.1.13). However, the fact that probability sets don't work well in situations where the choice depends on something not expressed in the model doesn't mean that you can't use a probability set to model such a situation, it just means that you shouldn't do it. This is what the following example shows.

Example 5.1.14.

Dropping the assumption $\mathsf{YI} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{C}|\mathsf{D}$ means that, in general, one or both of $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ or $\mathbb{P}_C^{\mathsf{Y}|\mathsf{ID}}$ may be ill-defined (note that the independence is merely a sufficient condition, not a necessary condition for these uniform conditional probabilities). The condition $\mathsf{YI} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{C}|\mathsf{D}$ alone also does *not* imply the conclusion of Theorem 5.1.9.

Constructing the following example requires the hypotheses that any given identifier $i \in \mathbb{N}$ could be associated with one of two input-output maps $D \rightarrow Y$. Thus the space of hypotheses is a sequence of binary values $H = \{0,1\}^{\mathbb{N}}$. Equipped with the product topology, H is a countable product of separable, completely metrizable spaces and is therefore also separable and completely metrizable (Willard, 1970, Thm. 16.4,Thm. 24.11). Thus $(H, \mathcal{B}(H))$ is a standard measurable space and, because it is uncountable, it is isomorphic to $([0,1], \mathcal{B}([0,1]))$.

Example 5.1.15. Take $Y = C = D = \{0,1\}$ and take $(H, \mathcal{H})$ to be $\{0,1\}^{\mathbb{N}}$ equipped with the product topology. For any $i \neq 1$, $\mathsf{Y}_i\mathsf{I}_i\mathsf{D}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{C}$, while $\mathbb{P}_\alpha^{\mathsf{D}_1} = \delta_\alpha$ and $\mathsf{I}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{C}$.

$\mathsf{YI} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{C}|\mathsf{D}$ follows from the fact that C can be (almost surely) written as a function of D.

For all $i, \in \mathbb{N}$, $y, d \in \{0,1\}$, $h \in H$ set

$$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{HI}_i\mathsf{D}_i}(y|h,j,d) = \delta_1(p(j,h))\delta_d(y) + \delta_0(p(j,h))\delta_{1-d}(y) \tag{5.50}$$

where $p(j,h)$ projects the j-th component of h. That is, if h maps j to 1, Y goes with D while if h maps j to 0, Y goes opposite D. Suppose also

$$\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{X}_{<i}, \mathsf{Y}_{<i}, \mathsf{I}_{<i}, \mathsf{C})|(\mathsf{X}_i, \mathsf{Y}_i, \mathsf{H}) \tag{5.51}$$

Then $\mathbb{P}_C^{\mathsf{Y}|\mathsf{DI}}$ is causally contractible. Set $\mathbb{P}_C^{\mathsf{H}}$ to be the uniform measure on $(H, \mathcal{H})$ and for $i > 1$

$$\mathbb{P}_C^{\mathsf{D}_i|\mathsf{I}_i\mathsf{H}}(d|j,h) = \delta_{p(j,h)}(d) \tag{5.52}$$

that is, if h maps j to 1, D is 1 while if h maps j to 0, D is 0. This also implies

$$\mathbb{P}_C^{\mathsf{I}_i|\mathsf{D}_i\mathsf{H}}(p(\cdot,h)^{-1}(d)|d,h) = 1 \tag{5.53}$$

Then, for $i > 1$

$$\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{HD}_i}(y|h,d) = \sum_{j\in\mathbb{N}} \delta_1(p(j,h))\delta_d(y)\mathbb{P}_C^{\mathsf{I}_i|\mathsf{D}_i\mathsf{H}}(j|d,h) + \delta_0(p(j,h))\delta_{1-d}(y)\mathbb{P}_C^{\mathsf{I}_i|\mathsf{D}_i\mathsf{H}}(j|d,h) \tag{5.54}$$

$$= \sum_{j\in\mathbb{N}} \delta_1(d)\delta_d(y)\mathbb{P}_C^{\mathsf{I}_i|\mathsf{D}_i\mathsf{H}}(j|d,h) + \delta_0(d)\delta_{1-d}(y)\mathbb{P}_C^{\mathsf{I}_i|\mathsf{D}_i\mathsf{H}}(j|d,h) \qquad \text{by Eq 5.53} \tag{5.55}$$

$$= \delta_1(y) \tag{5.56}$$

$$\implies \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i}(y|d) = \delta_1(y) \tag{5.57}$$

For $q \in I$, set

$$\mathbb{P}_C^{\mathsf{I}|\mathsf{H}}(q|h) = \begin{cases} 0.5 & q = (1,2,3,4,...) \text{ or } (1,3,2,4,...) \\ 0 & \text{otherwise} \end{cases} \tag{5.58}$$

and set

$$\mathbb{P}_C^{\mathsf{H}|\mathsf{D}}(h) = \begin{cases} 0.5 & h = (0,1,0,1,1,...) \text{ or } h = (0,0,1,1,1,...) \\ 0 & \text{otherwise} \end{cases} \tag{5.59}$$

Let $\overline{H}$ be the support of $\mathbb{P}_C^{\mathsf{H}|\mathsf{D}}(h)$.

Then for $i = 1$

$$\mathbb{P}_\alpha^{\mathsf{Y}_1|\mathsf{D}_1}(y|h,d) = \sum_{h\in H}\sum_{j\in\mathbb{N}} \mathbb{P}_\alpha^{\mathsf{I}_1|\mathsf{D}_1\mathsf{H}}(j|d,h)\mathbb{P}_C^{\mathsf{H}|\mathsf{D}_1}(h|d)\left(\delta_1(p(j,h))\delta_d(y) + \delta_0(p(j,h))\delta_{1-d}(y)\right) \tag{5.60}$$

$$= \sum_{h\in\overline{H}} 0.5(\delta_1(p(1,h))\delta_d(y) + \delta_0(p(1,h))\delta_{1-d}(y)) \tag{5.61}$$

$$= \delta_{1-d}(y)) \tag{5.62}$$

$$\neq \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i}(y|h,d) \qquad i \neq 1 \tag{5.63}$$

Thus $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ is not causally contractible by Theorem 4.3.7.

However, given any finite permutation $\rho : \mathbb{N} \to \mathbb{N}$

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{I}}(y|q) = \sum_{h\in\overline{H}}\sum_{d\in\{0,1\}^{\mathbb{N}}}\prod_{i\in\mathbb{N}} \mathbb{P}_C^{\mathsf{Y}_i|\mathsf{I}_i\mathsf{D}_i\mathsf{H}}(y_i|q_i,d_i,h)\mathbb{P}_\alpha^{\mathsf{D}_i|\mathsf{I}_i\mathsf{H}}(d_i|q_i,h)\mathbb{P}_C^{\mathsf{H}}(h) \tag{5.64}$$

$$= \delta_{1-\alpha}(y_1)\delta_{(1)_{i\in\mathbb{N}}}(y_{>1}) \tag{5.65}$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{I}}(y|\rho^{-1}(q)) \tag{5.66}$$

$$= \mathbb{F}_\rho\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{I}}(y|q) \tag{5.67}$$

References

Holy Bible : Contemporary English Version. New York : American Bible Society, [1995] ľ1995, 1995. URL `https://search.library.wisc.edu/catalog/999953290302121`.

David J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, December 1981. ISSN 0047-259X. doi: 10.1016/0047-259X(81)90099-3. URL `https://www.sciencedirect.com/science/article/pii/0047259X81900993`.

A. V. Banerjee, S. Chassang, and E. Snowberg. Chapter 4 - Decision Theoretic Approaches to Experiment Design and External ValidityaaWe thank Esther

Duflo for her leadership on the handbookǎand 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.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On Pearl's hierarchy and the foundations of causal inference. Technical report, 2020. URL https://causalai.net/r60.pdf.

Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction, 1998. URL http://jmvidal.cse.sc.edu/lib/sutton98a.html.

David A. Blackwell. *Theory of Games and Statistical Decisions*. Dover Publications, New York, September 1979. ISBN 978-0-486-63831-7.

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

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

Ethan D. Bolker. A Simultaneous Axiomatization of Utility and Subjective Probability. *Philosophy of Science*, 34(4):333–340, 1967. ISSN 0031-8248. URL https://www.jstor.org/stable/186122. Publisher: [The University of Chicago Press, Philosophy of Science Association].

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.

G. Chiribella, Giacomo D'Ariano, and P. Perinotti. Quantum Circuit Architecture. *Physical review letters*, 101:060401, September 2008. doi: 10.1103/PhysRevLett.101.060401.

Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, August 2019. ISSN 0960-1295, 1469-8072. doi: 10.1017/S0960129518000488.

Panayiota Constantinou and A. Philip Dawid. Extended conditional independence and applications in causal inference. *The Annals of Statistics*, 45(6):2618–2653, 2017. ISSN 0090-5364. URL http://www.jstor.org/stable/26362953. Publisher: Institute of Mathematical Statistics.

A. P. Dawid. Causal Inference without Counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424, June 2000. ISSN 0162-1459. doi: 10.1080/01621459.2000.10474210. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.2000.10474210. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.2000.10474210.

A. Philip Dawid. Decision-theoretic foundations for statistical causality. *arXiv:2004.12493 [math, stat]*, April 2020. URL http://arxiv.org/abs/2004.12493. arXiv: 2004.12493.

Bruno de Finetti. Foresight: Its Logical Laws, Its Subjective Sources. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Foundations and Basic Theory*, Springer Series in Statistics, pages 134–174. Springer, New York, NY, [1937] 1992. ISBN 978-1-4612-0919-5. doi: 10.1007/978-1-4612-0919-5_10. URL https://doi.org/10.1007/978-1-4612-0919-5_10.

P. Diaconis. Recent progress on de Finettis notions of exchangeability. *Bayesian Statistics*, 3:111–125, 1988.

P. Diaconis and D. Freedman. Finite Exchangeable Sequences. *The Annals of Probability*, 8(4):745–764, August 1980. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1176994663. URL https://projecteuclid.org/journals/annals-of-probability/volume-8/issue-4/Finite-Exchangeable-Sequences/10.1214/aop/1176994663.full. Publisher: Institute of Mathematical Statistics.

Dean Eckles and Eytan Bakshy. Bias and High-Dimensional Adjustment in Observational Studies of Peer Effects. *Journal of the American Statistical Association*, 116(534):507–517, April 2021. ISSN 0162-1459. doi: 10.1080/01621459.2020.1796393. URL https://doi.org/10.1080/01621459.2020.1796393. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2020.1796393.

Thomas S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, July 1967. ISBN 978-1-4832-2123-6.

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

Ronald A. Fisher. Cancer and Smoking. *Nature*, 182(4635):596–596, August 1958. ISSN 1476-4687. doi: 10.1038/182596a0. URL https://www.nature.com/articles/182596a0. Number: 4635 Publisher: Nature Publishing Group.

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.

J. Dmitri Gallow. The Causal Decision Theorist's Guide to Managing the News. *The Journal of Philosophy*, 117(3):117–149, March 2020. doi: 10.5840/jphil202011739. URL https://www.pdcnet.org/pdc/bvdb.nsf/purchase?openform&fp=jphil&id=jphil_2020_0117_0003_0117_0149.

Brett R. Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook. SSRN Scholarly Paper ID 3033144, Social Science Research Network, Rochester, NY, September 2018. URL https://papers.ssrn.com/abstract=3033144.

Brett R. Gordon, Robert Moakler, and Florian Zettelmeyer. Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement. *arXiv:2201.07055 [econ]*, January 2022. URL http://arxiv.org/abs/2201.07055. arXiv: 2201.07055.

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.

J. Y. Halpern. A Counter Example to Theorems of Cox and Fine. *Journal of Artificial Intelligence Research*, 10:67–85, February 1999. ISSN 1076-9757. doi: 10.1613/jair.536. URL https://www.jair.org/index.php/jair/article/view/10223.

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. Beyond exchangeability: The other conditions for causal inference in medical research. *Statistical Methods in Medical Research*, 21(1): 3–5, February 2012. ISSN 0962-2802. doi: 10.1177/0962280211398037. URL https://doi.org/10.1177/0962280211398037. Publisher: SAGE Publications Ltd STM.

Miguel A. Hernán and James M Robins. *Causal Inference: What If.* Chapman & Hall/CRC, 2020. URL https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

Eric Horvitz, David Heckerman, and Curtis Langlotz. A Framework for Comparing Alternative Formalisms for Plausible Reasoning. January 1986. URL https://openreview.net/forum?id=rJNeX0gdbr.

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.

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

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

Olav Kallenberg. The Basic Symmetries. In *Probabilistic Symmetries and Invariance Principles*, Probability and Its Applications, pages 24–68. Springer, New York, NY, 2005a. 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.

Olav Kallenberg. *Probabilistic Symmetries and Invariance Principles.* Probability and Its Applications. Springer-Verlag, New York, 2005b. ISBN 978-0-387-25115-8. doi: 10.1007/0-387-28861-9. URL http://link.springer.com/10.1007/0-387-28861-9.

Maximilian Kasy. Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead. *Political Analysis*, 24(3):324–338, 2016. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpw012. URL https://www.cambridge.org/core/journals/political-analysis/article/abs/why-experimenters-might-not-always-want-to-randomize-and-what-they-could-do-instead/79A12E7AB0921C50ECC1F1EA61EC1F8D. Publisher: Cambridge University Press.

G. Jay. Kerns and Gábor J. Székely. Definettis Theorem for Abstract Finite Exchangeable Sequences. *Journal of Theoretical Probability*, 19(3):589–608, December 2006. ISSN 1572-9230. doi: 10.1007/s10959-006-0028-z. URL https://doi.org/10.1007/s10959-006-0028-z.

Chayakrit Krittanawong, Bharat Narasimhan, Zhen Wang, Joshua Hahn, Hafeez Ul Hassan Virk, Ann M. Farrell, HongJu Zhang, and WH Wilson Tang. Association between chocolate consumption and risk of coronary artery disease: a systematic review and meta-analysis:. *European Journal of Preventive Cardiology*, July 2020. doi: 10.1177/2047487320936787. URL http://journals.sagepub.com/doi/10.1177/2047487320936787. Publisher: SAGE PublicationsSage UK: London, England.

David Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59(1): 5–30, March 1981. ISSN 0004-8402. doi: 10.1080/00048408112340011. URL https://doi.org/10.1080/00048408112340011.

D. V. Lindley and Melvin R. Novick. The Role of Exchangeability in Inference. *The Annals of Statistics*, 9(1):45–58, 1981. ISSN 0090-5364. URL `https://www.jstor.org/stable/2240868`. Publisher: Institute of Mathematical Statistics.

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

Naomi Oreskes and Erik M. Conway. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Climate Change: How a Handful of Scientists ... Issues from Tobacco Smoke to Global Warming*. Bloomsbury Press, New York, NY, June 2011. ISBN 978-1-60819-394-3.

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

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

Robert N. Proctor. The history of the discovery of the cigarettelung cancer link: evidentiary traditions, corporate denial, global toll. *Tobacco Control*, 21(2):87–91, March 2012. ISSN 0964-4563, 1468-3318. doi: 10.1136/tobaccocontrol-2011-050338. URL `https://tobaccocontrol.bmj.com/content/21/2/87`. Publisher: BMJ Publishing Group Ltd Section: The shameful past.

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.

Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov Properties for Acyclic Directed Mixed Graphs. *arXiv:1701.06686 [stat]*, January 2017. URL `http://arxiv.org/abs/1701.06686`. arXiv: 1701.06686.

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`.

Olli Saarela, David A. Stephens, and Erica E. M. Moodie. The role of exchangeability in causal inference. June 2020. doi: 10.48550/arXiv.2006.01799. URL `https://arxiv.org/abs/2006.01799v3`.

L. J. Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46:55–67, 1951. ISSN 1537-274X(Electronic),0162-1459(Print). doi: 10.2307/2280094.

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.

Ilya Shpitser and Judea Pearl. Complete Identification Methods for the Causal Hierarchy. *Journal of Machine Learning Research*, 9(Sep):1941–1979, 2008. ISSN ISSN 1533-7928. URL https://www.jmlr.org/papers/v9/shpitser08a.html.

Brian Skyrms. Causal Decision Theory. *The Journal of Philosophy*, 79(11):695–711, November 1982. doi: 10.2307/2026547. URL https://www.pdcnet.org/pdc/bvdb.nsf/purchase?openform&fp=jphil&id=jphil_1982_0079_0011_0695_0711.

Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

Statista. Cigarettes - worldwide | Statista Market Forecast, 2020. URL https://www.statista.com/outlook/50010000/100/cigarettes/worldwide.

Katie Steele and H. Orri Stefánsson. Decision Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2020 edition, 2020. URL https://plato.stanford.edu/archives/win2020/entries/decision-theory/.

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

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.

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

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

Robert Wiblin. Why smoking in the developing world is an enormous problem and how you can help save lives, 2016. URL https://80000hours.org/problem-profiles/tobacco/.

Stephen Willard. *General topology*. Reading, Mass., Addison-Wesley Pub. Co, 1970. ISBN 978-0-201-08707-9. URL http://archive.org/details/generaltopology00will_0.

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/.

World Health Organisation. Tobacco Fact sheet no 339, 2018. URL https://www.webcitation.org/6gUXrCDKA.

Appendix: