

# Causal Statistical Decision Theory|What are interventions?

David Johnston

May 6, 2022



# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Theories of causal inference . . . . .	5
<b>2</b>	<b>Technical Prerequisites</b>	<b>9</b>
2.1	Conventions . . . . .	10
2.2	Probability Theory . . . . .	10
2.2.1	Standard Probability Theory . . . . .	10
2.3	String Diagrams . . . . .	15
2.3.1	Elements of string diagrams . . . . .	15
2.3.2	Special maps . . . . .	17
2.3.3	Commutative comonoid axioms . . . . .	18
2.3.4	Manipulating String Diagrams . . . . .	18
2.4	Probability Sets . . . . .	20
2.4.1	Almost sure equality . . . . .	22
2.4.2	Extended conditional independence . . . . .	23
2.4.3	Examples . . . . .	25
2.4.4	Maximal probability sets and valid conditionals . . . . .	27
2.4.5	Existence of conditional probabilities . . . . .	31
<b>3</b>	<b>Models with choices and consequences</b>	<b>35</b>
3.1	What is the point of causal inference? . . . . .	36
3.1.1	Modelling decision problems . . . . .	37
3.1.2	Formal definitions . . . . .	38
3.2	Representation theorems for decision problems . . . . .	39
3.2.1	von Neumann-Morgenstern utility . . . . .	40
3.2.2	Savage decision theory . . . . .	40
3.2.3	Jeffrey's decision theory . . . . .	43
3.2.4	Causal decision theory . . . . .	44
3.2.5	Statistical decision theory . . . . .	45
3.3	Variables . . . . .	53
3.3.1	Variables and measurement procedures . . . . .	53
3.3.2	Measurement procedures . . . . .	54
3.3.3	Observable variables . . . . .	56
3.3.4	Model variables . . . . .	56

3.3.5	Variable sequences and partial order . . . . .	57
3.3.6	Decision procedures . . . . .	57
<b>4</b>	<b>Decision problems with repeatable phenomena</b>	<b>59</b>
4.1	Relevance to previous work . . . . .	60
4.2	Repeatable Response Functions . . . . .	61
4.2.1	Causally contractible Markov kernels . . . . .	62
4.2.2	Causal contractibility with data-independent actions . . .	72
4.3	Causal contractibility in sequences of active choices . . . . .	78
4.4	Response conditionals with data-dependent actions . . . . .	86
4.4.1	Combs . . . . .	87
4.4.2	Response conditionals in models with data dependent actions	88
4.4.3	Combs are the output of the “fix” operation . . . . .	90
4.5	Weaker assumptions than causal contractibility . . . . .	91
<b>5</b>	<b>See-do models, interventions and counterfactuals</b>	<b>95</b>
5.1	How do see-do models relate to other approaches to causal inference?	95
5.2	Interpretations of the choice set . . . . .	95
5.3	Causal Bayesian Networks as see-do models . . . . .	95
5.4	Unit Potential Outcomes models . . . . .	96
5.4.1	D-causation . . . . .	98
5.4.2	D-causation vs Limited Unresponsiveness . . . . .	100
5.4.3	Properties of D-causation . . . . .	103
5.4.4	Decision sequences and parallel decisions . . . . .	103
5.5	Existence of counterfactuals . . . . .	103
<b>6</b>	<b>Other causal modelling frameworks</b>	<b>107</b>

# 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 laboratory 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  $[X_i, X_j]$ -ideal intervention is an operation whose result is determined by applying the *do-calculus* to a *causally sufficient* model  $((\Omega, \mathcal{F}, \mathbb{P}), \mathcal{G}, \mathbf{U})$
- A model  $((\Omega, \mathcal{F}, \mathbb{P}), \mathcal{G}, \mathbf{U})$  is  $[X_i, X_j]$ -causally sufficient if  $\mathbf{U}$  contains  $X_i, X_j$  and “all intervenable variables that *cause*” both  $X_i$  and  $X_j$ <sup>1</sup>

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.

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

<sup>1</sup>Weaker conditions for causal sufficiency are possible, but they don’t avoid circularity (Shpitser and Pearl, 2008)

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



## 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  $\mathbf{X} : \Omega \rightarrow X$  and  $\mathbf{Y} : \Omega \rightarrow Y$ ,  $\mathbb{P}_C^{\mathbf{Y}|\mathbf{X}}$  represents a Markov kernel  $X \rightarrow Y$  that satisfies the definition of the distribution of  $\mathbf{Y}$  given  $\mathbf{X}$  (Definition 2.2.16) for every  $\alpha \in C$ , while  $\mathbb{P}_\alpha^{\mathbf{Y}|\mathbf{X}}$  is a conditional distribution only for  $\alpha$ . There are two unusual feature: firstly, it is more common to write a conditional distribution  $\mathbb{P}(\mathbf{Y}|\mathbf{X})$  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  $\mathbb{P}(\mathbf{Y}|\mathbf{X})$  is an aesthetic preference; writing a conditional distribution like this suggests  $\mathbb{P}(\mathbf{Y}|\mathbf{X})$  is the result of function composition between  $\mathbb{P}$  and some function denoted “ $\mathbf{Y}|\mathbf{X}$ ”. However, conditional probabilities are not given by composition of functions like this.

Name	notation	meaning
Iverson bracket	$\llbracket \cdot \rrbracket$	Function equal to 1 if $\cdot$ is true, false otherwise
Identity function	$\text{idf}_X$	Identity function $X \rightarrow X$
Identity kernel	$\text{id}_X$	Kernel associated with the identity function $X \rightarrow X$

## 2.2 Probability Theory

### 2.2.1 Standard Probability Theory

#### $\sigma$ -algebras

**Definition 2.2.1** (Sigma algebra). Given a set  $A$ , a  $\sigma$ -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  $\{B_i | i \in \mathbb{N}\}$  of elements of  $\mathcal{A}$ ,  $\cup_{i \in \mathbb{N}} B_i \in \mathcal{A}$

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

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

**Common  $\sigma$  algebras** For any  $A$ ,  $\{\emptyset, A\}$  is a  $\sigma$ -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  $\sigma$ -algebra.

Given  $A$  and a collection of subsets of  $B \subset \mathcal{P}(A)$ ,  $\sigma(B)$  is the smallest  $\sigma$ -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  $\sigma$ -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  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  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  $\mu : \mathcal{E} \rightarrow [0, 1]$  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  $\mathbb{M} : E \times \mathcal{F} \rightarrow [0, 1]$  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  $\mathbb{M} : E \times \mathcal{F} \rightarrow [0, 1]$ , we write the signature of  $\mathbb{M} : E \rightarrow F$ , read “ $\mathbb{M}$  maps from  $E$  to probability measures on  $F$ ”.

**Definition 2.2.8** (Deterministic Markov kernel). A *deterministic* Markov kernel  $\mathbb{A} : E \rightarrow \Delta(\mathcal{F})$  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  $\delta_x(A) = \mathbb{I}[x \in A]$

**Definition 2.2.10** (Markov kernel associated with a function). Given measurable  $f : (X, \mathcal{X}) \rightarrow (Y, \mathcal{Y})$ ,  $\mathbb{F}_f : X \rightarrow Y$  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  $\mathbb{Q}_\mu : \{*\} \rightarrow X$  is the unique Markov kernel  $* \mapsto \mu$

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

*Proof.*

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

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

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

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

□

### 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  $X : (\Omega, \mathcal{F}) \rightarrow (X, \mathcal{X})$ .

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

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

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

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

$$\iff \quad (2.6)$$

$$\mu^{XY} = \left( \triangleleft^{\mu^X} \right) \bullet \left( \mu^{Y|X} \right) \rightarrow Y \quad (2.7)$$

### 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.

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

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

$$\mu\mathbb{M}(A) := \int_X \mathbb{M}(A|x)\mu(dx) \quad (2.8)$$

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

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

$$\mathbb{M}\mathbb{N}(A|x) := \int_Y \mathbb{N}(A|y)\mathbb{M}(dy|x) \quad (2.9)$$

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

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

$$\mathbb{M}f(x) := \int_Y f(y)\mathbb{N}(dy|x) \quad (2.10)$$

for all  $x \in X$ .

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

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

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  $\mathbf{X} : \Omega \rightarrow (X, \mathcal{X})$ , define  $\mathbb{F}_{\mathbf{X}} : \Omega \rightarrow X$  by  $\mathbb{F}_{\mathbf{X}}(A|\omega) = \delta_{\mathbf{X}(\omega)}(A)$ , then*

$$\mu^{\mathbf{X}} = \mu\mathbb{F}_{\mathbf{X}} \quad (2.12)$$

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

$$\mu\mathbb{F}_{\mathbf{X}}(A) = \int_{\Omega} \delta_{\mathbf{X}(\omega)}(A) d\mu(\omega) \quad (2.13)$$

$$= \int_{\mathbf{X}^{-1}(\omega)} d\mu(\omega) \quad (2.14)$$

$$= \mu^{\mathbf{X}}(A) \quad (2.15)$$

□

### Semidirect product

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

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

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

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

$$(\mathbb{K} \odot \mathbb{L}) \odot \mathbb{M} = \mathbb{K} \odot (\mathbb{L} \odot \mathbb{M}) \quad (2.17)$$

$$(2.18)$$

*Proof.*

$$(\mathbb{K} \odot \mathbb{L}) \odot \mathbb{M} = \begin{array}{c} \text{Diagram showing the composition of kernels } \mathbb{K}, \mathbb{L}, \text{ and } \mathbb{M} \text{ for } (\mathbb{K} \odot \mathbb{L}) \odot \mathbb{M}. \end{array} \quad (2.19)$$

$$= \begin{array}{c} \text{Diagram showing the composition of kernels } \mathbb{K}, \mathbb{L}, \text{ and } \mathbb{M} \text{ for } \mathbb{K} \odot (\mathbb{L} \odot \mathbb{M}). \end{array} \quad (2.20)$$

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

□

The semidirect product can be used to define a notion of almost sure equality: two kernels  $\mathbb{K} : X \rightarrow Y$  and  $\mathbb{L} : X \rightarrow Y$  are  $\mu$ -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,  $\mathbb{K} \stackrel{\mu}{\cong} \mathbb{L}$  if and only if  $\mathbb{K} = \mathbb{L}$   $\mu$ -almost everywhere.

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

$$\mu \odot \mathbb{K} = \mu \odot \mathbb{L} \quad (2.22)$$

**Theorem 2.2.25.** Given  $(\mu, X, \mathcal{X})$ ,  $\mathbb{K} : X \rightarrow Y$  and  $\mathbb{L} : X \rightarrow Y$ ,  $\mathbb{K} \stackrel{\mu}{\cong} \mathbb{L}$  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.  $\square$

We often want to talk about almost sure equality of two different versions  $\mathbb{K}$  and  $\mathbb{L}$  of a conditional distribution  $\mathbb{P}^{Y|X}$  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  $\mathbb{P}^X$ . The relevant variables are usually obvious from the context and we leave them implicit and we will write  $\mathbb{K} \stackrel{\mathbb{P}}{\cong} \mathbb{L}$ . If the relevant marginal is ambiguous, we will instead write  $\mathbb{K} \stackrel{\mathbb{P}^X}{\cong} \mathbb{L}$ .

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

## 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:

$$\mathbb{K} := \boxed{\text{---} \mathbb{K} \text{---}} \quad (2.23)$$

$$\mu := \triangleleft \text{---} \mathbb{P} \text{---} \quad (2.24)$$

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

$$\mathbb{L}\mathbb{M} := X \boxed{\mathbb{K}} \boxed{\mathbb{M}} Z \quad (2.25)$$

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

$$\mathbb{K} \otimes \mathbb{L} := \begin{array}{c} W \boxed{\mathbb{K}} Y \\ X \boxed{\mathbb{L}} Z \end{array} \quad (2.26)$$

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

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

$$= \begin{array}{c} X \text{---} \bullet \boxed{\mathbb{K}} \text{---} \bullet \boxed{\mathbb{L}} \text{---} Z \\ \text{---} \bullet \boxed{\mathbb{L}} \text{---} Y \end{array} \quad (2.28)$$

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

$$\text{Id}^X := X \text{-----} X \quad (2.29)$$

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  $X \times Y$ :

$$X \times Y \cong \text{Id}^X \otimes \text{Id}^Y := \begin{array}{c} X \text{---} X \\ Y \text{---} Y \end{array} \quad (2.30)$$

$$= X \times Y \text{-----} X \times Y \quad (2.31)$$

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

$$X \text{---} \boxed{\mathbb{L}} \text{---} \begin{array}{c} Y \\ Z \end{array} \quad (2.32)$$

$$\equiv \quad (2.33)$$

$$X \text{---} \boxed{\mathbb{L}} \text{---} Y \times Z \quad (2.34)$$

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

$$\text{id}_X := X \text{---} X \quad (2.35)$$

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

$$\text{del}_X := \text{---} * X \quad (2.36)$$

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

$$\text{swap}_{X,Y} := \begin{array}{c} \diagup \quad \diagdown \\ \diagdown \quad \diagup \end{array} \quad (2.37)$$

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

$$\text{copy}_X := X \text{---} \begin{array}{c} \diagup \\ \diagdown \end{array} \begin{array}{c} X \\ X \end{array} \quad (2.38)$$

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

$$\text{copy}_X^1 = \text{copy}_X \quad (2.39)$$

$$\text{copy}_X^n = \begin{array}{c} \text{---} \boxed{\text{copy}_X^{n-1}} \text{---} \\ \bullet \diagdown \end{array} \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \end{array} \quad n > 1 \quad (2.40)$$

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

$$\bigotimes_{i \in A} \mathbb{K}_i := \boxed{\text{---} \boxed{\mathbb{K}_i} \text{---}}_{i \in A} \text{copy}_X^{|A|} \left( \bigotimes_{i \in A} \mathbb{K}_i \right) := \text{---} \bullet \boxed{\text{---} \boxed{\mathbb{K}_i} \text{---}}_{i \in A} \quad (2.41)$$

### 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  $\mathbb{K} : X \rightarrow Y$

$$(2.48)$$

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

$$\begin{array}{c} \text{---} \\ \text{---} \end{array} \begin{array}{c} \boxed{\mathbb{K}} \\ \boxed{\mathbb{L}} \end{array} \text{---} = (\text{id}_X \otimes \mathbb{L})\mathbb{K} \quad (2.49)$$

$$[(\text{id}_X \otimes \mathbb{L})\mathbb{K}](A|x, w) = \int_Y \int_X \mathbb{K}(A|x', y') \mathbb{L}(dy'|w) \delta_x(dx') \quad (2.50)$$

$$= \int_Y \mathbb{K}(A|x, y') \mathbb{L}(dy'|w) \quad (2.51)$$

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

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

$$\begin{array}{c} \text{---} \\ \text{---} \end{array} \bullet \begin{array}{c} \boxed{\mathbb{K}} \\ \bullet \end{array} \text{---} = (\text{id}_X \otimes \text{copy}_Y)\mathbb{K} \quad (2.52)$$

$$[(\text{id}_X \otimes \text{copy}_Y)\mathbb{K}](A|x, y) = \int_Y \int_Y \mathbb{K}(A|x, y', y'') \delta_y(dy') \delta_y(dy'') \quad (2.53)$$

$$= \mathbb{K}(A|x, y, y) \quad (2.54)$$

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

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

$$\begin{array}{c} \text{---} \\ \text{---} \end{array} \begin{array}{c} \boxed{\mathbb{K}} \\ \text{---} \end{array} = \text{swap}_{YX}\mathbb{K} \quad (2.55)$$

$$(\text{swap}_{YX}\mathbb{K})(A|y, x) = \int_{X \times Y} \mathbb{K}(A|x', y') \delta_y(dy') \delta_x(dx') \quad (2.56)$$

$$= \mathbb{K}(A|x, y) \quad (2.57)$$

The swap map before a kernel switches the input arguments.

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

$$\text{---} \begin{array}{c} \boxed{\mathbb{K}} \\ \text{---} \end{array} \begin{array}{c} \text{---} \\ \text{---} \end{array} = \mathbb{K}\text{swap}_{YZ} \quad (2.58)$$

$$(\mathbb{K}\text{swap}_{YZ})(A \times B|x) = \int_{Y \times Z} \delta_y(B) \delta_z(A) \mathbb{K}(dy \times dz|x) \quad (2.59)$$

$$= \int_{B \times A} \mathbb{K}(dy \times dz|x) \quad (2.60)$$

$$= \mathbb{K}(B \times A|x) \quad (2.61)$$

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

$$(\mathbb{K} \odot \mathbb{L})(\text{id}_Y \otimes \text{del}_Z) = \begin{array}{c} X \text{ --- } \boxed{\mathbb{K}} \text{ --- } \bullet \begin{array}{l} \text{--- } Y \\ \text{--- } \boxed{\mathbb{L}} \text{ --- } * \end{array} \end{array} \quad (2.62)$$

$$= \begin{array}{c} X \text{ --- } \boxed{\mathbb{K}} \text{ --- } \bullet \begin{array}{l} \text{--- } Y \\ \text{--- } * \end{array} \end{array} \quad \text{by Eq. 2.47} \quad (2.63)$$

$$= X \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \quad \text{by Eq. 2.43} \quad (2.64)$$

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

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

$$= \int_A \mathbb{K}(\text{d}y|x) \quad (2.66)$$

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

## 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  $X : \Omega \rightarrow X$  and a probability set  $\mathbb{P}_C$ , the

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

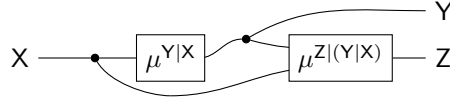
**Definition 2.4.3** (Uniform conditional distribution). Given a sample space  $(\Omega, \mathcal{F})$ , variables  $X : \Omega \rightarrow X$  and  $Y : \Omega \rightarrow Y$  and a probability set  $\mathbb{P}_C$ , a uniform conditional distribution  $\mathbb{P}_C^{Y|X}$  is any Markov kernel  $X \rightarrow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

Given a conditional distribution  $\mu^{ZY|X}$  we can define a higher order conditional  $\mu^{Z|(Y|X)}$ , which is a version of  $\mu^{Z|XY}$ . This is useful because uniform conditionals don't always exist, but we can use higher order conditionals to show that if a probability set  $\mathbb{P}_C$  has a uniform conditional  $\mathbb{P}_C^{ZY|X}$  then it also has a uniform conditional  $\mathbb{P}_C^{Z|XY}$  (Theorems 2.4.30 and 2.4.32). Given  $\mu^{XY|Z}$  and  $X : \Omega \rightarrow X$ ,  $Y : \Omega \rightarrow Y$  standard measurable, it has recently been proven that a higher order conditional  $\mu^{Z|(Y|X)}$  exists Bogachev and Malofeev (2020), Theorem 3.5.

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

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

$$\iff \quad (2.69)$$



$$\mu^{ZY|X} = \quad (2.70)$$

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

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

$$\int_A \mathbb{K}(B|x) \mathbb{P}_\alpha^X(dx) = \int_A \mathbb{L}(B|x) \mathbb{P}_\alpha^X(dx) \quad (2.71)$$

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

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

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

$$\mathbb{P}_\alpha^X \odot \mathbb{K} = \mathbb{P}_\alpha^X \odot \mathbb{L} \quad (2.72)$$

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

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

$$\mathbb{P}_\alpha^X \odot \mathbb{K} = \mathbb{P}_\alpha^{XY} \quad (2.73)$$

$$= \mathbb{P}_\alpha^X \odot \mathbb{L} \quad (2.74)$$

□

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

$$\mathbb{P}_\alpha^{Y|X} \odot \mathbb{K} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_\alpha^{Y|X} \odot \mathbb{L} \quad (2.75)$$

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

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

$$\stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_\alpha^X \odot (\mathbb{P}_C^{Y|X} \odot \mathbb{K}) \quad (2.77)$$

$$\stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_\alpha^X \odot (\mathbb{P}_C^{Y|X} \odot \mathbb{L}) \quad (2.78)$$

□

**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  $X : \Omega \rightarrow X$ ,  $Y : \Omega \rightarrow Y$  and uniform conditional distributions  $\mathbb{P}_C^{Y|X}$  and  $\mathbb{P}_C^{Z|XY}$ , then  $\mathbb{P}_C^{YZ|X}$  exists and is equal to*

$$\mathbb{P}_C^{YZ|X} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y|X} \odot \mathbb{P}_C^{Z|XY} \quad (2.79)$$

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

$$\mathbb{P}_\alpha^{XYZ} = \mathbb{P}_\alpha^X \odot \mathbb{P}_\alpha^{YZ|X} \quad (2.80)$$

$$= \mathbb{P}_\alpha^X \odot (\mathbb{P}_\alpha^{Y|X} \odot \mathbb{P}_\alpha^{Z|YX}) \quad (2.81)$$

$$= \mathbb{P}_\alpha^X \odot (\mathbb{P}_C^{Y|X} \odot \mathbb{P}_C^{Z|YX}) \quad (2.82)$$

□

### 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  $A, B, Z$ , we say  $B$  is conditionally independent of  $A$  given  $C$ , written  $B \perp\!\!\!\perp_{\mathbb{P}_\alpha} A|C$ , if

$$\mathbb{P}^{Y|WX} \stackrel{\mathbb{P}}{\cong} \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (2.83)$$

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

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  $W : \Omega \rightarrow W$ ,  $X : \Omega \rightarrow X$ ,  $Y : \Omega \rightarrow Y$ ,  $Y \perp\!\!\!\perp_{\mathbb{P}} X|W$  if and only if there exists some version of  $\mathbb{P}^{Y|WX}$  and  $\mathbb{K} : W \rightarrow Y$  such that*

$$\mathbb{P}^{Y|WX} \stackrel{\mathbb{P}}{\cong} \begin{array}{c} W \text{ --- } \boxed{\mathbb{K}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (2.85)$$

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

*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  $\Omega$

**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  $\phi : C \rightarrow X$ .

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

**Definition 2.4.14** (Complementary nonstochastic variables). A pair of nonstochastic variables  $\phi$  and  $\xi$  are complementary if  $(\phi, \xi)$  is invertible.

**Example 2.4.15.** The letters  $\phi$  and  $\xi$  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  $\phi$  and  $\xi$ , the extended conditional independence  $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e X \phi | Z \xi$  holds if for each  $a \in \xi(C)$ ,  $\mathbb{P}_{\xi^{-1}(a)}^{Y|XZ}$  and  $\mathbb{P}_{\xi^{-1}(a)}^{Y|X}$  exist and

$$\mathbb{P}_{\xi^{-1}(a)}^{Y|XZ} \stackrel{\mathbb{P}_{\xi^{-1}(a)}}{\cong} \begin{array}{c} Z \text{ --- } \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (2.87)$$

$$\iff \quad (2.88)$$

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

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

$$\mathbb{P}_C^{Y|XZ} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} Z \text{ --- } \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ X \text{ --- } * \end{array} \quad (2.90)$$

$$\iff \quad (2.91)$$

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

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:  $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Y\phi|Z\xi$  iff  $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e X\phi|Z\xi$
2.  $X \perp\!\!\!\perp_{\mathbb{P}_C}^e YC|YC$
3. Decomposition:  $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Y\phi|W\xi$  and  $Z \preceq Y$  implies  $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Z\phi|W\xi$
4. Weak union:
  - (a)  $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Y\phi|W\xi$  and  $Z \preceq Y$  implies  $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Y\phi|(Z, W)\xi$
  - (b)  $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Y\phi|W\xi$  and  $\lambda \preceq \phi$  implies  $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Y\phi|(Z, W)(\xi, \lambda)$
5. Contraction:  $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Z\phi|W\xi$  and  $X \perp\!\!\!\perp_{\mathbb{P}_C}^e Y\phi|(Z, W)\xi$  implies  $X \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y, Z)\phi|W\xi$

The following forms of these properties are often used here:

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

### 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  $(D, D \circ \mathcal{S}_\alpha)$  where  $D \circ \mathcal{S}_\alpha$  always yields  $\alpha$ .

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

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

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

$$\mathbb{P}_\alpha^{Y|D} = \mathbb{P}_\alpha^Y \quad (2.94)$$

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

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

Because only deterministic marginals  $\mathbb{P}_\alpha^D$  are available, for every  $\alpha \in C$  we have  $Y \perp\!\!\!\perp_{\mathbb{P}_\alpha}^e D$ . This reflects the fact that *after we have selected a choice*  $\alpha$  the value of  $C$  provides no further information about the distribution of  $Y$ , because  $D$  is deterministic given any  $\alpha$ . 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  $\mathbb{P}_C^{XY|Z}$ ,*

$$\mathbb{P}_C^{XY|Z} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \quad (2.96)$$

$$\iff \quad (2.97)$$

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

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

*Proof.* If: By Theorem 2.4.32

$$\mathbb{P}_C^{XY|Z} = \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|ZX}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \quad (2.99)$$

$$\stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \quad (2.100)$$

$$= \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \quad (2.101)$$

Only if: Suppose

$$\mathbb{P}_C^{XY|Z} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} Z \text{ --- } \bullet \begin{cases} \boxed{\mathbb{P}_C^{Y|Z}} \text{ --- } Y \\ \boxed{\mathbb{P}_C^{X|Z}} \text{ --- } X \end{cases} \end{array} \quad (2.102)$$

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

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

then

$$\mathbb{P}_\alpha^{XYZZ}(A \times B \times C) = \int_{A \times C} \mathbb{P}_C^{Y|XZ}(B|x, z) \mathbb{P}_C^{X|Z}(dx|z) \mathbb{P}_\alpha^Z(dz) \quad (2.104)$$

$$> \int_{A \times C} \mathbb{P}_C^{Y|X}(B|z) \mathbb{P}_C^{X|Z}(dx|z) \mathbb{P}_\alpha^Z(dz) \quad (2.105)$$

$$= \int_C \mathbb{P}_C^{XY|X}(A \times B|z) \mathbb{P}_\alpha^Z(dz) \quad (2.106)$$

$$= \mathbb{P}_\alpha^{XYZZ}(A \times B \times C) \quad (2.107)$$

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( $X = x$ )” specify a probability set by a collection of uniform conditional probabilities on variables other than  $X$ , along with marginal probabilities of  $X$ . Specifically:

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

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,  $2X$  is not equal to  $X$ , but we cannot define a causal Bayesian network where both  $X$  and  $2X$  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  $X \longrightarrow Y$  and some observational probability distribution  $\mathbb{P}^{XY} \in \Delta(X \times Y)$ . Using the standard notion of “hard interventions on  $X$ ”, this model induces a probability set which we could informally describe as the set  $\mathbb{P}_{\square} := \{\mathbb{P}_a^{XY} | a \in X \cup \{*\}\}$  where  $*$  is a special element corresponding to the observational setting. The graph  $X \longrightarrow Y$  implies the existence of the uniform conditional probability  $\mathbb{P}_{\square}^{Y|X}$  under the nominated set of interventions, while the usual rules of hard interventions imply that  $\mathbb{P}_a^X = \delta_a$  for  $a \in X$ .

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

Consider, for example,  $\Omega = \{0, 1\}$  with  $X = (Z, Z)$  for  $Z := \text{id}_{\Omega}$  and any measure  $\kappa \in \Delta(\{0, 1\}^2)$  such that  $\kappa(\{1\} \times \{0\}) > 0$ . Note that  $X^{-1}(\{1\} \times \{0\}) =$

$Z^{-1}(\{1\}) \cap Z^{-1}(\{0\}) = \emptyset$ . Thus for any probability measure  $\mu \in \Delta(\{0, 1\})$ ,  $\mu^{\mathbf{X}}(\{1\} \times \{0\}) = \mu(\emptyset) = 0$  and so  $\kappa$  cannot be the marginal distribution of  $\mathbf{X}$  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  $\mathbf{X} : \Omega \rightarrow X$ , an  $\mathbf{X}$ -valid probability distribution is any probability measure  $\mathbb{K} \in \Delta(X)$  such that  $\mathbf{X}^{-1}(A) = \emptyset \implies \mathbb{K}(A) = 0$  for all  $A \in \mathcal{X}$ .

**Definition 2.4.21** (Valid conditional). Given  $(\Omega, \mathcal{F})$ ,  $\mathbf{X} : \Omega \rightarrow X$ ,  $\mathbf{Y} : \Omega \rightarrow Y$  a  $\mathbf{Y}|\mathbf{X}$ -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:

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

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

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

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

$$\forall A \in \mathcal{X} : \mathbf{X} \bowtie A = \emptyset \implies \mathbb{P}^{\mathbf{X}|\ast}(A|\ast) = 0 \quad (2.110)$$

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

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

□

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

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  $X_i|X_{[i-1]}$ -valid conditionals  $\{\mathbb{P}_i^{X_i|X_{[i-1]}} | i \in [n]\}$ ; then recursively taking the semidirect product  $\mathbb{M} := \mathbb{P}_1^{X_1} \odot (\mathbb{P}_2^{X_2|X_1} \odot \dots)$  yields a  $X_{[n]}$  valid distribution. Furthermore, the maximal probability set associated with  $\mathbb{M}$  is nonempty.

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

Note that validity is not a necessary condition for a conditional to define a non-empty probability set. Given some  $\mathbb{K} : X \rightarrow Y$ ,  $\mathbb{K}$  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  $(\Omega, \mathcal{F})$ ,  $X : \Omega \rightarrow X$ ,  $Y : \Omega \rightarrow Y$ ,  $Z : \Omega \rightarrow Z$  (all spaces standard measurable) and any valid candidate conditional  $\mathbb{P}^{Y|X}$  and  $\mathbb{Q}^{Z|YX}$ ,  $\mathbb{P}^{Y|X} \odot \mathbb{Q}^{Z|YX}$  is also a valid candidate conditional.*

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

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

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

$$\mathbb{R}^{YZ|X}(B \times C|x) = \int_B \mathbb{Q}^{Z|YX}(C|y, x) \mathbb{P}^{Y|X}(dy|x) \quad (2.112)$$

$$\leq \mathbb{P}^{Y|X}(B|x) \quad (2.113)$$

$$= 0 \quad (2.114)$$

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

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

$$= 0 \quad (2.116)$$

□

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

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

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

*Proof.* If: this follows directly from Corollary 2.4.25.

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

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

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

$$> 0 \quad (2.119)$$

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

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

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

$$\mathbb{P}_{\square}^{B|WHC} = \begin{array}{c} H \text{ --- } * \\ C \text{ ----- } B \\ W \text{ --- } * \end{array} \quad (2.120)$$

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

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

$$= \emptyset \quad (2.122)$$

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

Validity rules out conditional probabilities like 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  $X : \Omega \rightarrow X$  and  $Y : \Omega \rightarrow Y$ ,  $Z : \Omega \rightarrow Z$  and a probability set  $\mathbb{P}_C$  with conditional  $\mathbb{P}_C^{X|Y}$  such that  $Z = f \circ Y$  for some  $f : Y \rightarrow Z$ . Then there exists a conditional probability  $\mathbb{P}_C^{Z|X} = \mathbb{P}_C^{Y|X} \mathbb{F}_f$ .*

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

$$\mathbb{P}_\alpha^{XZ} = \mathbb{P}_\alpha^{XY} \mathbb{F}_{\text{id}_X \otimes f} \quad (2.123)$$

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

$$\mathbb{F}_{\text{id}_X \otimes f}(A \times B | x, y) = \delta_x(A) \delta_{f(y)}(B) \quad (2.124)$$

$$= \mathbb{F}_{\text{id}_X}(A | x) \otimes \mathbb{F}_f(B | y) \quad (2.125)$$

$$\implies \mathbb{F}_{\text{id}_X \otimes f} = \mathbb{F}_{\text{id}_X} \otimes \mathbb{F}_f \quad (2.126)$$

Thus

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

$$= \begin{array}{c} \text{X} \\ \curvearrowright \\ \begin{array}{c} \triangleleft \mathbb{P}_\alpha^X \\ \bullet \\ \boxed{\mathbb{P}_C^{Y|X}} \end{array} \text{---} \boxed{\mathbb{F}_f} \text{---} Z \end{array} \quad (2.128)$$

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

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

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

**Theorem 2.4.30** (Existence of higher order valid conditionals with respect to probability sets). *Suppose we have a sample space  $(\Omega, \mathcal{F})$ , variables  $\mathbf{X} : \Omega \rightarrow X$  and  $\mathbf{Y} : \Omega \rightarrow Y$ ,  $\mathbf{Z} : \Omega \rightarrow Z$  and a probability set  $\mathbb{P}_C$  with regular conditional  $\mathbb{P}_C^{\mathbf{Y}\mathbf{Z}|\mathbf{X}}$  and  $Y$  and  $Z$  standard measurable. Then there exists a regular  $\mathbb{P}_C^{\mathbf{Z}|\mathbf{Y}|\mathbf{X}}$ .*

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

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

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

In particular,  $\mathbb{P}_C^{\mathbf{Y}\mathbf{Z}|\mathbf{X}}$  is a Borel measurable map  $X \rightarrow Y \times Z$ . Thus equation 2.130 implies for all  $A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z}$

$$\mathbb{P}_C^{\mathbf{Y}\mathbf{Z}|\mathbf{X}}(\mathbf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x, y) \mathbb{P}_C^{\mathbf{Y}\mathbf{Z}|\mathbf{X}} \mathbb{F}_f(dy|x) \quad (2.131)$$

$$= \int_A n(B|x, y) \mathbb{P}_C^{\mathbf{Y}|\mathbf{X}}(dy|x) \quad (2.132)$$

Where Equation 2.132 follows from Lemma 2.4.28.

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

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

which implies  $n$  is a version of  $\mathbb{P}_C^{\mathbf{Y}\mathbf{Z}|\mathbf{Y}|\mathbf{X}}$ . By Lemma 2.4.28,  $n\mathbb{F}_f$  is a version of  $\mathbb{P}_C^{\mathbf{Z}|\mathbf{Y}|\mathbf{X}}$ .  $\square$

We might be motivated to ask whether the higher order conditionals in Theorem 2.4.30 can be chosen to be valid. Despite Lemma 2.4.31 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 2.4.31.** *Given a probability space  $(\mu, \Omega, \mathcal{F})$  and variables  $\mathbf{X} : \Omega \rightarrow X$ ,  $\mathbf{Y} : \Omega \rightarrow Y$ , if there is a regular proper conditional probability  $\mu^{|\mathbf{X}} : X \rightarrow \Omega$  then there is a valid conditional distribution  $\mu^{\mathbf{Y}|\mathbf{X}}$ .*



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

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

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

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

$$= 0 \quad (2.135)$$

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

**Theorem 2.4.32** (Higher order conditionals). *Suppose we have a sample space  $(\Omega, \mathcal{F})$ , variables  $\mathbf{X} : \Omega \rightarrow X$  and  $\mathbf{Y} : \Omega \rightarrow Y$ ,  $\mathbf{Z} : \Omega \rightarrow Z$  and a probability set  $\mathbb{P}_C$  with conditional  $\mathbb{P}_C^{\mathbf{YZ}|\mathbf{X}}$ . Then  $\mathbb{P}_C^{\mathbf{Z}|\mathbf{Y}|\mathbf{X}}$  is a version of  $\mathbb{P}_C^{\mathbf{Z}|\mathbf{Y}\mathbf{X}}$*

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

$$\mathbb{P}_{\alpha}^{\mathbf{YZ}|\mathbf{X}} = \quad (2.136)$$

$$\Rightarrow \mathbb{P}_{\alpha}^{\mathbf{XYZ}} = \quad (2.137)$$

$$= \quad (2.138)$$

$$= \quad (2.139)$$

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

**Theorem 2.4.33.** *Given probability gap model  $\mathbb{P}_C$ ,  $\mathbf{X}$ ,  $\mathbf{Y}$ ,  $\mathbf{Z}$  such that  $\mathbb{P}_C^{\mathbf{Z}|\mathbf{Y}\mathbf{X}}$  exists,  $\mathbb{P}_C^{\mathbf{Z}|\mathbf{Y}}$  exists iff  $\mathbf{Z} \perp_{\mathbb{P}_C} \mathbf{X}|\mathbf{Y}$ .*

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

$$\mathbb{P}_{\alpha}^{\mathbf{Y}|\mathbf{WX}} = \quad (2.140)$$

□

**Theorem 2.4.34** (Valid higher order conditionals). *Suppose we have a sample space  $(\Omega, \mathcal{F})$ , variables  $\mathbf{X} : \Omega \rightarrow X$  and  $\mathbf{Y} : \Omega \rightarrow Y$ ,  $\mathbf{Z} : \Omega \rightarrow Z$  and a probability set  $\mathbb{P}_C$  with regular conditional  $\mathbb{P}_C^{\mathbf{Y}\mathbf{Z}|\mathbf{X}}$ ,  $Y$  discrete and  $Z$  standard measurable. Then there exists a valid regular  $\mathbb{P}_C^{\mathbf{Z}|\mathbf{X}\mathbf{Y}}$ .*

*Proof.* By Theorem 2.4.30, we have a higher order conditional  $\mathbb{P}_C^{\mathbf{Z}|\mathbf{Y}|\mathbf{X}}$  which, by Theorem 2.4.32 is also a version of  $\mathbb{P}_C^{\mathbf{Z}|\mathbf{X}\mathbf{Y}}$ .

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

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

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

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

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

## 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 6 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 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 6 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  $n$ th 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  $\Omega$  along with a utility function  $u : \Omega \rightarrow \mathbb{R}$

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

- A Markov kernel  $C \rightarrow \Omega$
- A Markov kernel  $C \times H \rightarrow \Omega$  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  $\Omega \rightarrow \mathbb{R}$  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  $\Omega$  and a utility function  $u : F \rightarrow \mathbb{R}$ . We call these ingredients a “decision problem”.

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

Our task is to find a *model* that relates  $C$  to  $\Omega$ . 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  $(C, \Omega, u)$ , a *choices only model* is a function  $C \rightarrow \Omega$ .

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

Both types of models induce probability sets.

**Definition 3.1.4** (Induced probability set). Given a decision problem  $(C, \Omega, u)$  and a model  $\mathbb{P} : C \times H \rightarrow \Omega$ , 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  $C \rightarrow \Delta(\Omega)$  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  $C \rightarrow \Delta(\Omega)$  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  $u : \Omega \rightarrow \mathbb{R}$  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 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*  $\Omega$  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  $\Omega \rightarrow \mathbb{R}$ . This is much closer to the subject of this work than the theorem of von Neumann and Morgenstern.

**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  $\Omega$  and a set of acts  $C$  such that  $|C| = \Omega^S$  and an evaluation function  $T : S \times C \rightarrow F$  such that for any  $f : S \rightarrow \Omega$  there exists  $c \in C$  such that  $T(\cdot, c) = f$ .

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

$$\alpha \preceq \alpha' \iff \int_S u(T(s, \alpha)) \mu(ds) \leq \int_S u(T(s, \alpha')) \mu(ds) \quad \forall \alpha, \alpha' \in C \quad (3.1)$$

*Proof.* Savage (1954) □

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



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

$$\alpha \preceq \alpha' \iff \int_{\Omega} u(f) \mathbb{P}_{\alpha}(df) \leq \int_{\Omega} u(f) \mathbb{P}_{\alpha'}(df) \quad \forall \alpha, \alpha' \in C \quad (3.2)$$

*Proof.* Define  $\mathbb{P} : C \rightarrow \Delta(\Omega)$  by

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

where  $T_{\alpha} : S \rightarrow F$  is the function  $s \mapsto T(s, \alpha)$ .  $\mathbb{P}_{\alpha}$  is the pushforward of  $T_{\alpha}$  under  $\mu$ .

Then

$$\int_{\Omega} u(f) \mathbb{P}_{\alpha}(df) = \int_S u \circ T_{\alpha}(s) \mu(ds) \quad (3.4)$$

$$= \int_S u(T(s, \alpha)) \mu(ds) \quad (3.5)$$

□

### Savage axioms

Careful analysis of Savage's theorem is outside the scope of this work, but given the relevance 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  $T : S \times C \rightarrow F$ , let all greek letters  $\alpha, \beta$  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  $\alpha'$  and  $\beta'$  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  $\alpha, \beta$  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  $A, B \in \mathcal{S}$ ,  $A \leq B$  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  $\leq$  as “is less probable than”
- P4: For every  $A, B \in \mathcal{S}$ ,  $A \leq B$  or  $B \leq A$
- P5: For some  $\alpha, \beta$ ,  $\alpha \prec \beta$
- P6: Suppose  $\alpha \not\leq \beta$ . Then for every  $\gamma$  there is a finite partition of  $S$  such that if  $\alpha'$  agrees with  $\alpha$  and  $\beta'$  agrees with  $\beta$  except on some element  $B$  of the partition,  $\alpha'$  and  $\beta'$  being equal to  $\gamma$  on  $B$ , then  $\alpha \not\leq \beta'$  and  $\alpha' \not\leq \beta$
- 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  $\Omega$  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  $S \rightarrow \Omega$ . 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  $\alpha$  being not preferred to another  $\beta$  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  $\alpha, \beta$  if  $\alpha$  is better than  $\beta$  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”  $\leq$  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  $\gamma$  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  $\Omega$ , 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 Savage's; where identifies a set of acts  $C$  with all functions  $S \rightarrow F$  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  $\text{des} : \underline{\mathcal{F}} \rightarrow \mathbb{R}$  and a probability distribution  $\mu \in \Delta(\underline{\mathcal{F}})$  such that for  $A, B \in \underline{\mathcal{F}}$  and finite partition  $D_1, \dots, D_n \in \underline{\mathcal{F}}$ :*

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

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  $C \rightarrow \Delta(\underline{\mathcal{F}})$  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  $\mathbb{P} : C \rightarrow \Delta(\underline{\mathcal{F}})$  such that for any  $\alpha, \alpha' \in C$  and finite partition  $D_1, \dots, D_n \in \underline{\mathcal{F}}$ :*

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

Where  $\mu$  and  $\text{des}$  are as in Theorem 3.2.4

*Proof.* Define  $\mathbb{P}$ . 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  $\sigma$ -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  $G \not\prec A$ , 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  $G \not\prec A$  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  $\Omega$  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  $H : \Omega \rightarrow H$  (the *dependency hypothesis*) and  $S : \Omega \rightarrow S$  (the *consequence*) and a utility  $u : \Omega \rightarrow \mathbb{R}$ , the *causal utility* of a choice  $\alpha \in C$  is given by

$$U(\alpha) := \int_S \int_H u(s) \mathbb{P}_\alpha^{S|H}(ds|h) \mathbb{P}_C^H(dh) \quad (3.8)$$

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

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  $\mathbb{P} : C \rightarrow \Delta(\Omega)$ . 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 6 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,  $l : D \times H \rightarrow \mathbb{R}$  is a loss function and  $\mathbb{P} : H \rightarrow \mathcal{X}$  is a Markov kernel from hypotheses 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 requires 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  $R : C \times H \rightarrow \mathbb{R}$  is defined by

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

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  $R : C \times H \rightarrow \mathbb{R}$  and prior  $\mu \in \Delta(H)$ , the  $\mu$ -Bayes risk is

$$R_\mu(\mathbb{D}_\alpha) := \int_H R(\mathbb{D}_\alpha, h) \mu(dh) \quad (3.10)$$

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

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

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

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

### 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 6. 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  $(\mathbb{P}_{C \times H}, \mathbf{X}, \mathbf{Y}, \mathbf{D})$  where  $\mathbb{P}_{C \times H}$  is a probability set indexed by elements of  $C \times H$  on  $(\Omega, \mathcal{F})$ ,  $\mathbf{X} : \Omega \rightarrow X$  are the observations,  $\mathbf{Y} : \Omega \rightarrow Y$  are the consequences and  $\mathbf{D} : \Omega \rightarrow D$  are the decisions.  $\mathbb{P}_{C \times H}$  must observe the following conditional independences:

$$\mathbf{X} \perp\!\!\!\perp_{\mathbb{P}_{C \times H}}^e \mathbf{C} | \mathbf{H} \quad (3.13)$$

$$\mathbf{D} \perp\!\!\!\perp_{\mathbb{P}_{C \times H}}^e \mathbf{H} | \mathbf{C} \quad (3.14)$$

where  $\mathbf{C} : C \times H \rightarrow C$  and  $\mathbf{H} : C \times H \rightarrow H$  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  $(\mathbb{P}_{C \times H}, \mathbf{X}, \mathbf{Y}, \mathbf{D})$  where the following additional conditional independence holds:

$$\mathbf{Y} \perp\!\!\!\perp_{\mathbb{P}_{C \times H}}^e (\mathbf{X}, \mathbf{C}) | (\mathbf{D}, \mathbf{H}) \quad (3.15)$$

We assume that a utility function is available depending on the consequence  $\mathbf{Y}$  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  $(\mathbb{P}_{C \times H}, \mathbf{X}, \mathbf{Y}, \mathbf{D})$  and a utility  $u : Y \rightarrow \mathbb{R}$ , the induced loss  $l : D \times H \rightarrow \mathbb{R}$  is defined as

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

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

A see-do model induces a set of decision functions: for each  $\alpha \in C$ , there is an associated probability distribution  $\mathbb{P}_{\alpha}^{\mathbf{D}|\mathbf{X}}$ . 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  $(\mathbb{P}_{C \times H}, \mathbf{X}, \mathbf{Y}, \mathbf{D})$  along with a utility  $u : Y \rightarrow \mathbb{R}$ , 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  $\mathbb{P}_{\alpha}^{\mathbf{D}|\mathbf{X}}$  and hypothesis  $h$ .*

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

*Proof.* The expected utility given  $\alpha$  and  $h$  is

$$\int_Y u(y) \mathbb{P}_{\alpha, h}^{\mathbf{Y}}(dy) = \int_Y \int_D \int_X u(y) \mathbb{P}_{\alpha, h}^{\mathbf{Y}|\mathbf{D}\mathbf{X}}(dy|d, x) \mathbb{P}_{\alpha, h}^{\mathbf{D}|\mathbf{X}}(dd|x) \mathbb{P}_{\alpha, h}^{\mathbf{X}}(dx) \quad (3.18)$$

$$= \int_X \int_D \int_Y u(y) \mathbb{P}_{\alpha, h}^{\mathbf{Y}|\mathbf{D}}(dy|d) \mathbb{P}_{\alpha, h}^{\mathbf{D}|\mathbf{X}}(dd|x) \mathbb{P}_{\alpha, h}^{\mathbf{X}}(dx) \quad (3.19)$$

$$= \int_X \int_D \int_Y u(y) \mathbb{P}_{C \times \{h\}}^{\mathbf{Y}|\mathbf{D}}(dy|d) \mathbb{P}_{\{\alpha\} \times H}^{\mathbf{D}|\mathbf{X}}(dd|x) \mathbb{P}_{C \times \{h\}}^{\mathbf{X}}(dx) \quad (3.20)$$

$$= - \int_D \int_X l(d, h) \mathbb{P}_{\{\alpha\} \times H}^{\mathbf{D}|\mathbf{X}}(dd|x) \mathbb{P}_{C \times \{h\}}^{\mathbf{X}}(dx) \quad (3.21)$$

$$= -R(\mathbb{P}_{\{\alpha\} \times H}^{\mathbf{D}|\mathbf{X}}, h) \quad (3.22)$$

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

Theorem 3.2.16 does *not* hold for general see-do models. General see-do models allow for the utility to depend on  $\mathbf{X}$  even after conditioning on  $\mathbf{D}$  and  $\mathbf{H}$ , 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  $\mathbf{Y}$ .



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

$$R(\alpha, h) := -\mathbb{P}_{\alpha, h}^{\mathbf{X}\mathbf{Y}} u \quad \forall \alpha \in C, h \in H \quad (3.23)$$

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  $(\mathbb{P}_{C \times H}, \mathbf{X}, \mathbf{Y}, \mathbf{D})$  and a prior over hypotheses  $\mu \in \Delta(H)$  can be 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  $(\mathbb{P}_C, \mathbf{X}, \mathbf{Y}, \mathbf{D}, \mathbf{H})$  where  $\mathbb{P}_C$  is a probability set on  $(\Omega, \mathcal{F})$ ,  $\mathbf{X} : \Omega \rightarrow X$  are the observations,  $\mathbf{Y} : \Omega \rightarrow Y$  are the consequences,  $\mathbf{D} : \Omega \rightarrow D$  are the decisions and  $\mathbf{H} : \Omega \rightarrow H$  is the hypothesis.  $\mathbb{P}_C$  must observe the following conditional independences:

$$\mathbf{X} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{C} | \mathbf{H} \quad (3.24)$$

$$\mathbf{D} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{H} | \mathbf{C} \quad (3.25)$$

$$\mathbf{H} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{C} \quad (3.26)$$

**Definition 3.2.19** (Induced Bayesian see-do model). Given a see-do model  $(\mathbb{P}_{C \times H}, \mathbf{X}, \mathbf{Y}, \mathbf{D}, \mathbf{H})$  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

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

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

**Theorem 3.2.20** (Induced SDP Bayes risk). *Given a conditionally independent see-do model  $(\mathbb{P}_C, \mathbf{X}, \mathbf{Y}, \mathbf{D}, \mathbf{H})$  along with a utility  $u : Y \rightarrow \mathbb{R}$  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  $\mu$ -Bayes risk of that decision rule.*

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

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

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

$$\mathbb{P}_\alpha^Y u = \int_Y u(y) \mathbb{P}_\alpha^Y(dy) \quad (3.28)$$

$$= \int_Y \int_D \int_X \int_H u(y) \mathbb{P}_\alpha^{Y|DXH}(dy|d, x, h) \mathbb{P}_\alpha^{D|XH}(dd|x, h) \mathbb{P}_\alpha^{X|H}(dx|h) \mathbb{P}_\alpha^H(dh) \quad (3.29)$$

$$= \int_X \int_D \int_Y \int_H u(y) \mathbb{P}_\alpha^{Y|DH}(dy|d, h) \mathbb{P}_\alpha^{D|X}(dd|x) \mathbb{P}_\alpha^{X|H}(dx|h) \mathbb{P}_\alpha^H(dh) \quad (3.30)$$

$$= \int_X \int_D \int_Y \int_H u(y) \mathbb{P}_C^{Y|DH}(dy|d, h) \mathbb{P}_\alpha^{D|X}(dd|x) \mathbb{P}_C^{X|H}(dx|h) \mu(dh) \quad (3.31)$$

$$= - \int_D \int_X \int_H l(d, h) \mathbb{P}_\alpha^{D|X}(dd|x) \mathbb{P}_C^{X|H}(dx|h) \mu(dh) \quad (3.32)$$

$$= - \int_H R(\mathbb{P}_\alpha^{D|X}, h) \mu(dh) \quad (3.33)$$

$$= -R_\mu(\mathbb{P}_\alpha^{D|X}) \quad (3.34)$$

□

### 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, then 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  $R : C \times H \rightarrow \mathbb{R}$  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  $R : H \times C \rightarrow \mathbb{R}$ .

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  $R : C \times H \rightarrow \mathbb{R}$ , a choice  $\alpha \in C$  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  $\alpha$ .

**Definition 3.2.23** (Complete class). A *complete class* is any  $B \subset C$  such that, for any  $\alpha' \notin B$  there is some  $\alpha \in B$  that dominates  $\alpha'$ . 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  $R : C \times H \rightarrow \mathbb{R}$ , the risk set is the subset of  $\mathbb{R}^{|H|}$  given by

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

**Theorem 3.2.26** (Complete class theorem). *Given a risk function  $R : C \times H \rightarrow \mathbb{R}$ , 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  $R : C \times H \rightarrow \mathbb{R}$  and an associated probability set  $\mathbb{P}_C$  on  $(\Omega, \mathcal{F})$ ,  $\Omega$  finite, if the function*

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

*is linear and*

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

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

*Proof.* By linearity of

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

we also have linearity of

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

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.  $\square$

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

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

*is linear.*

*Proof.* By definition,

$$R(\alpha, h) = -\mathbb{P}_{\alpha,h}^{\mathbf{X}\mathbf{Y}} u \quad (3.41)$$

$$= -\mathbb{P}_{C \times \{h\}}^{\mathbf{X}} \odot \mathbb{P}_{\alpha \times h}^{\mathbb{D}|\mathbf{X}} \odot \mathbb{P}_{C \times \{h\}}^{\mathbf{Y}|\mathbf{D}\mathbf{X}} u \quad (3.42)$$

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

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

$$-\mathbb{P}_{C \times \{h\}}^{\mathbf{X}} \odot \mathbb{P}_{\alpha \times h}^{\mathbb{D}|\mathbf{X}} \odot \mathbb{P}_{\alpha,h}^{\Omega|\mathbf{D}\mathbf{X}} u \quad (3.43)$$

where  $\alpha$  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.  $\square$

**Example 3.2.30.** Suppose we have a conditionally independent see-do model  $(\mathbb{P}_C, \mathbf{X}, \mathbf{Y}, \mathbf{D}, \mathbf{H})$  along with a bounded utility  $u : Y \rightarrow \mathbb{R}$  where  $H, D, X$  and  $Y$  are all finite, and  $\{\mathbb{P}_\alpha^{\mathbb{D}|\mathbf{X}} | \alpha \in C\}$  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”  $S_\alpha$ , one for each choice  $\alpha \in C$ .  $S_\alpha$  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  $S$  by applying a function to its result. The function  $X$  that, when applied to the result of  $S$ , 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  $F = MA$ . 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  $F = MA$ , it must be understood relative to some measurement procedure  $S$  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,  $F(s) = M(s)A(s)$  will hold.

A measurement procedure  $S$  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.  $S$  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  $S$  is not a purely mathematical thing, we cannot perform mathematical reasoning with  $S$  directly. It is much more practical to relegate  $S$  to the background, and reason in terms of the functions  $F$ ,  $M$  and  $A$ . However, even if we don’t talk about it much,  $S$  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  $\bigvee_{x \in A} \mathcal{B} \bowtie x$ .

**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  $f : B \rightarrow C$ , define the “composition”  $f \circ \mathcal{B}$  to be any procedure  $\mathcal{C}$  that yields  $f(x)$  whenever  $\mathcal{B}$  yields  $x$ . We can construct such a procedure by describing the steps: first, do  $\mathcal{B}$  and secondly, apply  $f$  to the value yielded by  $\mathcal{B}$ .

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

$$(g \circ f) \circ \mathcal{B} \text{ yields } x \iff \mathcal{B} \text{ yields } (g \circ f)^{-1}(x) \quad (3.44)$$

$$\iff \mathcal{B} \text{ yields } f^{-1}(g^{-1}(x)) \quad (3.45)$$

$$\iff f \circ \mathcal{B} \text{ yields } g^{-1}(x) \quad (3.46)$$

$$\iff g \circ (f \circ \mathcal{B}) \text{ yields } x \quad (3.47)$$

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,  $\Psi$  will generally be a finite set. We can therefore equip  $\Psi$  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  $(X \circ \mathcal{S}, X)$  where  $X : (\Psi, \mathcal{E}) \rightarrow (X, \mathcal{X})$  is a measurable function and  $\mathcal{X} := X \circ \mathcal{S}$  is the measurement procedure induced by  $X$  and  $\mathcal{S}$ .

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

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

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

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

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

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

### 3.3.4 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  $S : \Omega \rightarrow \Psi$  called the complete observable variable, and every observable variable  $(X' \circ S, X')$  is associated with the model variable  $X := X' \circ S$ .

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

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

### 3.3.5 Variable sequences and partial order

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

Define the partial order on variables  $\preceq$  where  $X \preceq Y$  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  $X : \Omega \rightarrow X$ ,  $Y : \Omega \rightarrow Y$ ,  $X \preceq Y$  if there is some  $f : Y \rightarrow X$  such that  $X = f \circ Y$ .

Clearly,  $X \preceq (X, Y)$  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  $\alpha$ .

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

However,  $\mathcal{S}$  is problematic to model. The model is often part of the process of choosing  $\alpha$ , and so a model of  $\mathcal{S}$  that involves the step “choose  $\alpha$ ” 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

# Decision problems with repeatable phenomena

Chapter 2 introduced probability sets as generic tools for causal modelling, while Chapter 3 examined how probability set models can be used to construct mathematical models of decision problems. In general, few assumptions were made about the structure of the models in question. Section 3.2.5 added some structure with *see-do* models, which featured variables representing observations and consequences, and non-stochastic variables representing decisions and hypotheses. Extended conditional independence properties defined the roles of these different variables in the model. However, 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, nor do observations need to form a sequence. Causal inference in practice often concerns the question of making choices in a context where observations and consequences are repeatable, and the subject of this chapter is to examine probability sets that model decision problems with this kind of repeatability.

Repeatability in classical statistical models is often expressed by the assumption of *exchangeability of observations*. This is the assumption that the measurement procedure produces a sequence of values that are “all alike” in the particular sense that any rearrangement of the sequence should be modeled with the same model (although note that exchangeability is only implied by this assumption if observations are modeled with a unique probability model, see Walley (1991, pg. 463)). An exchangeable probability distribution over a sequence of variables is a mixture of *independent and identically distributed* distributions. Models with choices generally cannot be represented by a mixture of identically distributed sequences, because different choices will usually mean different actions will be taken and different distributions will result. The appropriate generalisation of independent and identical distributions seems to be independent and identical response functions – that is, two elements of the sequence will have the same distribution over *outputs* given identical *inputs*.

Chapter 6 reviews how this is a typical assumption of causal modelling frameworks, and this chapter investigates when models of sequences with choices will feature independent and identical response functions.

The key result of this chapter is that a model of choices that results in a sequence of variables is representable as a mixture of independent and identical response functions is equivalent to the assumption of *causal contractibility*. Causal contractibility is defined in Section 4.2.1 and compared to prior work on similar questions. A representation theorem for causally contractible Markov kernels is proved in Section 4.2.1. It is applied to probability sets modelling “one-shot” choices with no dependence on prior data in Section 4.2.2, and generalised to “adaptive” choices where actions may depend on prior data in Section 4.4; the latter generalisation requires the notion of *combs*, introduced in Section 4.4.1.

Causal contractibility is a more complicated assumption than exchangeability, and this chapter also considers the problem of assessing causal contractibility. In particular, Theorem 4.3.4 shows that if there are “experimental subjects” that are exchangeable in a particular way, and the input variables and choices are both equally informative in the sense that given either the inputs or the choice nothing relevant can be gained by learning the other, then causal contractibility holds. To my knowledge, this is the only example of a theorem of its type that proves the existence of “causal effects” in models featuring choices. Somewhat similar theorems exist that address the identification of potential outcomes, but they prove different things (the identifiability of a certain kind of latent variable, which is not identical to the consequences of a choice), and provide different conditions. A particular difference is that the conditions for Theorem 4.3.4 concern only *observable* variables, and while randomisation of actions is sometimes an admissible condition, it is not a necessary one.

The study of causal inference is often concerned with situations where Theorem 4.3.4 does not apply. A problem of particular interest involves going from “passive observations” to “active interventions”. Section 4.5 considers how this problem can be represented using probability set models, and why simple assumptions that would render it solvable are rarely acceptable.

## 4.1 Relevance to 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). This work is only engaged shallowly with this literature, but the idea that symmetries of probabilistic models may imply representability as mixtures of “fixed but unknown” models is crucial, as is the basic result of De Finetti.

The second line of work is the study of exchangeability-like assumptions in

causal models in particular. Dawid (2020) defines *post-treatment exchangeability*, closely related to *exchange commutativity* (Definition 4.2.2), one of the two conditions constituting causal contractibility. Greenland and Robins (1986) discusses the assumption of “exchangeability of individuals” in a medical experiment that also suggests the key idea of exchange commutativity. Banerjee et al. (2017) also mention the condition that “subjects are exchangeable conditional on covariates, so that experiments identical up to a permutation of labels are equivalent from the perspective of the experimenter”, which is similarly suggestive of exchange commutativity. While both of these are suggestive of exchange commutativity, exchanging *individuals* is a transformation of measurement procedures, not an operation defined on a probability model, and so it does not automatically imply any particular properties of the model. Exchange commutativity, on the other hand, is a symmetry of Markov kernels which might be appealing in situations where measurement procedures with individuals exchanged are regarded as essentially the same. Rubin (2005) discusses the assumption of the exchangeability of potential outcomes.

The other component of causal contractibility is *consequence locality* (Definition 4.2.1). This is also suggested by existing work – in particular, the stable unit treatment distribution assumption (SUTDA) in Dawid (2020), and the stable unit treatment value assumption (SUTVA) in (Rubin, 2005): “(“SUTVA) comprises two sub-assumptions. First, it assumes that *there is no interference between units* (Cox 1958); that is, neither  $Y_i(1)$  nor  $Y_i(0)$  is affected by what action any other unit received. Second, it assumes that *there are no hidden versions of treatments*; no matter how unit  $i$  received treatment 1, the outcome that would be observed would be  $Y_i(1)$  and similarly for treatment 0.

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 Repeatable Response Functions

Start with a sequence of variable pairs  $(X_i, Y_i)_{i \in \mathbb{N}}$  where  $X_i$  is the  $i$ th “input” and  $Y_i$  is the corresponding “output”, each taking values in  $X$  and  $Y$  respectively. A “repeatable response function” is a probabilistic mapping  $X \rightarrow Y$  that is identical for all  $(X_i, Y_i)$  pairs. Repeatable response functions are in general be unknown, in which case (under a Bayesian model  $\mathbb{P}_C$ ), the response function is *not* the conditional  $\mathbb{P}_C^{Y_i|X_i}$  but rather the conditional  $\mathbb{P}_C^{Y_i|X_i H}$  where  $H$  is an unobserved hypothesis variable. “Repeatability” also means that the same response function can be obtained no matter which actions have already been taken. That is,  $Y_i$  is independent of previous input-output pairs when conditioned on  $H$  and  $X_i$ .

The result of this section is that the pairs in a sequence  $(X, Y) := (X_i, Y_i)_{i \in \mathbb{N}}$  modeled by  $\mathbb{P}_C$  are related by repeatable response functions if and only if there is a causally contractible *uniform comb*  $\mathbb{P}_C^{Y|X}$ . Combs are a generalisation of conditional probabilities and, and if we assume that actions are independent of previous data ( $X_i \perp\!\!\!\perp_{\mathbb{P}_C}^e Y_{<i} C | X_{<i}$ ), this reduces to the assumption that the

uniform conditional distribution  $\mathbb{P}_C^{Y|X}$  is causally contractible.

Because combs are unfamiliar, this section is structured so that the “data-independent actions” case is introduced first. Specifically, the representation theorem is proven for general Markov kernels in Section 4.2.1, and applied to models  $\mathbb{P}_C$  with data-independent actions in Section 4.2.2. Subsequently, combs are introduced in Section 4.4, and the general result applied to models with data-dependent actions. The following section, Section 4.3, discusses questions related to when the assumption of causal contractibility might be held to apply to a particular problem, as well as introducing “Ecclesiastes’ assumption”, a weaker assumption than causal contractibility which could be applied to problems where observations and active interventions are mixed.

### 4.2.1 Causally contractible Markov kernels

Here a representation theorem is proved for causally contractible Markov kernels. First, causal contractibility is defined, which is the conjunction of the sub-assumptions of *locality* and *exchange commutativity*.

The intuitive basis of the two sub-assumptions is easier to see for Markov kernels with just two input-output pairs. In that simplified case, exchange commutativity for two inputs and outputs is given by the following equality:

$$\begin{array}{c} D_1 \\ D_2 \end{array} \begin{array}{c} \diagup \\ \diagdown \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y_{\{1,2\}}|D_{\{1,2\}}}} \\ \diagdown \\ \diagup \end{array} \begin{array}{c} Y_1 \\ Y_2 \end{array} = \begin{array}{c} D_1 \\ D_2 \end{array} \begin{array}{c} \diagdown \\ \diagup \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y_{\{1,2\}}|D_{\{1,2\}}}} \\ \diagup \\ \diagdown \end{array} \begin{array}{c} Y_1 \\ Y_2 \end{array} \quad (4.1)$$

It expresses the idea that swapping the inputs is equivalent to swapping the outputs. Locality is given by the following pair of equalities:

$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y_{1,2}|X_{1,2}}} \\ \diagdown \\ \diagup \end{array} \begin{array}{c} Y_1 \\ * \end{array} = \begin{array}{c} X_1 \\ X_2 \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y_1|X_1}} \\ \diagdown \\ \diagup \end{array} \begin{array}{c} Y_1 \\ * \end{array} \quad (4.2)$$

$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y_{1,2}|X_{1,2}}} \\ \diagup \\ \diagdown \end{array} \begin{array}{c} * \\ Y_2 \end{array} = \begin{array}{c} X_1 \\ X_2 \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y_2|X_2}} \\ \diagup \\ \diagdown \end{array} \begin{array}{c} * \\ Y_2 \end{array} \quad (4.3)$$

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

#### Definition of causal contractibility

The general definitions follow.

**Definition 4.2.1** (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

$$\begin{array}{ccc} \begin{array}{c} X^n \\ X^\mathbb{N} \end{array} \begin{array}{c} \diagup \\ \diagdown \end{array} \boxed{\mathbb{K}} \begin{array}{c} \diagdown \\ \diagup \end{array} \begin{array}{c} Y^n \\ * \end{array} & \begin{array}{c} X^n \\ X^\mathbb{N} \end{array} \begin{array}{c} \diagup \\ \diagdown \end{array} \boxed{\mathbb{L}} \begin{array}{c} \diagdown \\ \diagup \end{array} \begin{array}{c} Y^n \\ * \end{array} \\ & = \end{array} \quad (4.4)$$

$$\iff \quad (4.5)$$

$$\mathbb{K}(\bigtimes_{i \in [n]} A_i \times Y^\mathbb{N} | x_{[n]}, x_{[n]^c}) = \mathbb{L}(\bigtimes_{i \in [n]} A_i | x_{[n]}) \quad (4.6)$$

**Definition 4.2.2** (Exchange commutativity). A Markov kernel  $\mathbb{K} : X^\mathbb{N} \rightarrow Y^\mathbb{N}$  *commutes with exchange* if for all finite permutations  $\rho : \mathbb{N} \rightarrow \mathbb{N}$ ,  $A_i \in \mathcal{Y}$ ,  $(x_{[n]}, x_{[n]^c}) \in \mathbb{N}$

$$\mathbb{K} \text{swap}_{\rho, Y} = \text{swap}_{\rho, X} \mathbb{K} \quad (4.7)$$

$$\iff \quad (4.8)$$

$$\mathbb{K}(\bigtimes_{i \in \mathbb{N}} A_{\rho(i)} | (x_i)_{i \in \mathbb{N}}) = \mathbb{K}(\bigtimes_{i \in \mathbb{N}} A_i | (x_{\rho(i)})_{i \in \mathbb{N}}) \quad (4.9)$$

Causal contractibility is the conjunction of both assumptions.

**Definition 4.2.3** (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.

### Properties of causally contractible Markov kernels

A causally contractible Markov kernel over a sequence of pairs treats all subsequences as equivalent in a particular way (Theorem 4.2.7, although the sense in which this implies equivalence of subsequences might be more obvious from the statement of Theorem 4.2.16 presented later). This feature is the motivation for the name *causal contractibility*. Both sub-assumptions are independent – Theorem 4.2.8 presents counterexamples.

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

Lemma 4.2.5 is used to extend finite sequences to infinite ones, and is used in a number of upcoming theorems.

**Lemma 4.2.5** (Infinitely extended kernels). *Given a collection of Markov kernels  $\mathbb{K}_i : X^i \rightarrow Y^i$  for all  $i \in \mathbb{N}$ , if we have for every  $j > i$*

$$\mathbb{K}_j(id_{X_i} \otimes del_{X_{j-i}}) = \mathbb{K}_i \otimes del_{X_{j-i}} \quad (4.10)$$

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

$$\mathbb{K}(id_{X_i} \otimes del_{X_{j-i}}) = \mathbb{K}_i \otimes del_{X_{j-i}} \quad (4.11)$$

*Proof.* Take any  $x \in X^{\mathbb{N}}$  and let  $x_{|m} \in X^m$  be the first  $m$  elements of  $x$ . By Equation 4.10, for any  $A_i \in \mathcal{Y}$ ,  $i \in [m]$

$$\mathbb{K}_n(\bigtimes_{i \in [m]} A_i \times Y^{n-m} | x_{|n}) = \mathbb{K}_m(\bigtimes_{i \in [m]} A_i | x_{|m}) \quad (4.12)$$

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

$$\mathbb{K}_n \text{swap}_{\rho}(\bigtimes_{i \in [m]} A_{\rho(i)} \times Y^{n-m} | x_{|n}) = \mathbb{K}_n(\bigtimes_{i \in [m]} A_i \times Y^{n-m} | x_{|n}) \quad (4.13)$$

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

$$\mathbb{Q}_d(\bigtimes_{i \in [n]} A_i \times Y^{\mathbb{N}}) = \mathbb{K}_n(\bigtimes_{i \in [n]} A_{\rho(i)} | d_{|n}) \quad (4.14)$$

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

$$\mathbb{Q}_d(\bigtimes_{i \in [n]} A_i \times Y^{\mathbb{N}}) \geq \mathbb{Q}_d(\bigtimes_{i \in [p]} A_i \times Y^{\mathbb{N}}) \quad (4.15)$$

$$\geq \mathbb{Q}_d(\bigtimes_{i \in \mathbb{N}} A_i) \quad (4.16)$$

so by the Monotone convergence theorem, the sequence  $\mathbb{Q}_d(\bigtimes_{i \in [n]} A_i)$  converges as  $n \rightarrow \infty$  to  $\mathbb{Q}_d(\bigtimes_{i \in \mathbb{N}} A_i)$ .  $d \mapsto \mathbb{Q}_d^{\mathbb{Z}^n}(\bigtimes_{i \in [n]} A_i)$  is measurable for all  $n$ ,  $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$  by Equation 4.14, and so  $d \mapsto \mathbb{Q}_d$  is also measurable.

Thus  $d \mapsto \mathbb{Q}_d$  is the desired  $\mathbb{P}_C^{Y^{\mathbb{N}} D^{\mathbb{N}}} : D^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ .  $\square$

Theorem 4.2.7 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 4.2.6** (Marginalising kernel). Given  $(X, \mathcal{X})$  and  $A \subset \mathbb{N}$ ,  $\text{marg}_A : X^{\mathbb{N}} \rightarrow X^A$  is the Markov kernel given by

$$\bigotimes_{i \in \mathbb{N}} \text{switch}_{A,i} \quad (4.17)$$

where

$$\text{switch}_A = \begin{cases} \text{id}_X & i \in A \\ \text{del}_X & i \notin A \end{cases} \quad (4.18)$$

**Theorem 4.2.7** (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*

$$\mathbb{K} \text{marg}_A = \text{swap}_{[n] \leftrightarrow A} \mathbb{L} \otimes \text{del}_{X^{\mathbb{N}}} \quad (4.19)$$

*Proof.* Only if: By exchange commutativity

$$\text{swap}_{[n] \leftrightarrow A} \mathbb{K} = \mathbb{K} \text{swap}_{[n] \leftrightarrow A} \quad (4.20)$$

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

$$\text{swap}_{[n] \leftrightarrow A} \mathbb{K} \text{swap}_{[n] \leftrightarrow A} = \mathbb{K} \quad (4.21)$$

so

$$\mathbb{K} \text{marg}_A = \text{swap}_{[n] \leftrightarrow A} \mathbb{K} \text{swap}_{[n] \leftrightarrow A} \text{marg}_A \quad (4.22)$$

$$= \text{swap}_{[n] \leftrightarrow A} \mathbb{K} \text{marg}_{[n]} \quad (4.23)$$

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

$$\mathbb{K} \text{marg}_{[n]} = \mathbb{K}(\text{id}_{[n]} \otimes \text{del}_{X^{\mathbb{N}}}) \quad (4.24)$$

$$= \mathbb{L} \otimes \text{del}_{X^{\mathbb{N}}} \quad (4.25)$$

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

$$\text{swap}_{A \leftrightarrow [n]} \mathbb{K} \text{marg}_A = \text{swap}_{A \leftrightarrow [n]} \mathbb{K} \text{swap}_{A \leftrightarrow [n]} (\text{id}_{[n]} \otimes \text{del}_{X^{\mathbb{N}}}) \quad (4.26)$$

$$= \mathbb{K} \text{marg}_{[n]} \quad (4.27)$$

$$= \mathbb{K}(\text{id}_{[n]} \otimes \text{del}_{X^{\mathbb{N}}}) \quad (4.28)$$

Then by Lemma 4.2.5

$$\text{swap}_{A \leftrightarrow [n]} \mathbb{K} \text{swap}_{A \leftrightarrow [n]} = \mathbb{K} \quad (4.29)$$

Consider an arbitrary finite permutation  $\rho : \mathbb{N} \rightarrow \mathbb{N}$ .  $\rho$  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  $\rho$  itself can be written as a composition of a sequence of transpositions. Thus for any finite  $\rho : \mathbb{N} \rightarrow \mathbb{N}$

$$\text{swap}_\rho \mathbb{K} \text{swap}_\rho = \mathbb{K} \quad (4.30)$$

□

Theorem 4.2.8 shows that neither locality nor exchange commutativity is implied by the other.

**Theorem 4.2.8.** *Exchange commutativity does not imply locality or vice versa.*

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

$$\mathbb{K}(y_1, y_2 | d_1, d_2) = \mathbb{I}[(y_1, y_2) = (d_1 + d_2, d_1 + d_2)] \quad (4.31)$$

then

$$\mathbb{K}(y_1, Y | d_1, d_2) = \mathbb{I}[y_1 = d_1 + d_2] \quad (4.32)$$

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  $\rho$  to be the unique nontrivial swap  $\{0, 1\} \rightarrow \{0, 1\}$

$$\text{swap}_{\rho, D} \mathbb{K}(y_1, y_2 | d_1, d_2) = \mathbb{K}(y_1, y_2 | d_2, d_1) \quad (4.33)$$

$$= \mathbb{I}[(y_1, y_2) = (d_2 + d_1, d_2 + d_1)] \quad (4.34)$$

$$= \mathbb{I}[(y_1, y_2) = (d_1 + d_2, d_1 + d_2)] \quad (4.35)$$

$$= \mathbb{I}[(y_2, y_1) = (d_1 + d_2, d_1 + d_2)] \quad (4.36)$$

$$= \mathbb{K} \text{swap}_{\rho, Y}(y_1, y_2 | d_1, d_2) \quad (4.37)$$

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

$$\mathbb{K}(y_1, y_2 | d_1, d_2) = \mathbb{I}[(y_1, y_2) = (0, 1)] \quad (4.38)$$

Then:

$$\mathbb{K}(y_1 | d_1, d_2) = \mathbb{I}[y_1 = 0] \quad (4.39)$$

$$= \mathbb{K}(y_1 | d_1) \quad (4.40)$$

$$\mathbb{K}(y_2 | d_1, d_2) = \mathbb{I}[y_2 = 1] \quad (4.41)$$

$$= \mathbb{K}(y_2 | d_2) \quad (4.42)$$

so  $\mathbb{K}$  satisfies locality.

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

$$\text{swap}_{\rho(\mathbf{D})} \mathbb{K}(y_1, y_2 | d_1, d_2) = \mathbb{K}(y_1, y_2 | d_2, d_1) \quad (4.43)$$

$$= \mathbb{I}(y_1, y_2) = (0, 1) \quad (4.44)$$

$$\neq \mathbb{I}(y_2, y_1) = (0, 1) \quad (4.45)$$

$$= \mathbb{K}_{\text{swap}_{\rho(\mathbf{D})}}(y_1, y_2 | d_1, d_2) \quad (4.46)$$

□

A model of the treatment of several patients who have already been examined might satisfy consequence locality but not exchange commutativity. Patient B's treatment could 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 economic interventions might satisfy exchange commutativity but not locality. If a government prints money to make exactly  $n$  payments of \$10 000 are made to a number of undistinguished recipients, the government cannot say much about the impact of who exactly receives the payment. However, the amount of inflation created by the payments depends on the number of payments made; making 100 such payments will have a negligible effect on inflation, while making payments to everyone in the country will have a substantial effect, and this will in turn affect the outcomes of the people who did or did not receive payment. Dawid (2000) offers the alternative example of herd immunity in vaccination campaigns as a situation where commutativity of exchange holds but locality does not.

Although locality seems to imply a lack of 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  $\mathbf{D}_1 = 0$ ), or flip a second coin use this as the input for step 2 (this is the choice  $\mathbf{D}_1 = 1$ ). At the second step, I may further choose to copy the provisional results ( $\mathbf{D}_2 = 0$ ) or invert them ( $\mathbf{D}_2 = 1$ ). Then

$$\mathbb{P}_S^{Y_1 | \mathbf{D}}(y_1 | d_1, d_2) = 0.5 \quad (4.47)$$

$$\mathbb{P}_S^{Y_2 | \mathbf{D}}(y_2 | d_1, d_2) = 0.5 \quad (4.48)$$

- 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 4.2.1
- Nevertheless, the choice at step 1 affects the result of step 2

### Representation theorems for causally contractible Markov kernels

Theorem 4.2.9 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 4.2.9 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 4.2.11).

**Theorem 4.2.9.** *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*

$$\mathbb{K} = \begin{array}{c} \triangle \mu \\ \text{---} X \text{---} \boxed{\mathbb{F}_{\text{ev}}} \text{---} Y \end{array} \quad (4.49)$$

$$\iff \quad (4.50)$$

$$\mathbb{K}(A|(x_i)_{i \in \mathbb{N}}) = \mu \Pi_{(x_i)_{i \in \mathbb{N}}}(A) \forall A \in \mathcal{Y}^{\mathbb{N}} \quad (4.51)$$

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

$$(y_{ji})_{j,i \in X \times \mathbb{N}} \mapsto (y_{d_i i})_{i \in \mathbb{N}} \quad (4.52)$$

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

$$\text{ev} : X^{\mathbb{N}} \times Y^{X \times \mathbb{N}} \rightarrow Y \quad (4.53)$$

$$((x_i)_{i \in \mathbb{N}}, (y_{ji})_{j,i \in X \times \mathbb{N}}) \mapsto (y_{x_i i})_{i \in \mathbb{N}} \quad (4.54)$$

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

Define

$$\mu\left(\bigtimes_{(i,j) \in X \times \mathbb{N}} A_{ij}\right) := \mathbb{K}\left(\bigtimes_{(i,j) \in X \times \mathbb{N}} A_{ij} | e\right) \forall A_{ij} \in \mathcal{Y} \quad (4.55)$$

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

$$\mathbb{Q} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}} \quad (4.56)$$

$$\mathbb{Q} := \begin{array}{c} \triangle \mu \\ \text{---} X \text{---} \boxed{\mathbb{F}_{\text{ev}}} \text{---} Y \end{array} \quad (4.57)$$

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  $\text{swap}_{A \leftrightarrow B} : \mathbb{N} \rightarrow \mathbb{N}$  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  $x \in X^{\mathbb{N}}$

$$\mathbb{Q}(C_A | x) = \mu(\text{ev}_x^{-1}(C_A)) \quad (4.58)$$

The argument of  $\mu$  is

$$\text{ev}_x^{-1}(C_A) = \{(y_{ji})_{j, i \in X \times \mathbb{N}} | (y_{x_i i})_{i \in \mathbb{N}} \in C_A\} \quad (4.59)$$

$$= \bigtimes_{i \in \mathbb{N}} \bigtimes_{j \in X} D_{ji} \quad (4.60)$$

where

$$D_{ji} = \begin{cases} C_i & (j, i) \in B \\ Y & \text{otherwise} \end{cases} \quad (4.61)$$

and so

$$\text{swap}_{A \leftrightarrow B}(\text{ev}_x^{-1}(C_A)) = C_A \quad (4.62)$$

Substituting Equation 4.62 into 4.58

$$\mathbb{Q}(C_A | x) = \mu \text{swap}_{A \leftrightarrow B}(C_A) \quad (4.63)$$

$$= \mathbb{K} \text{swap}_{A \leftrightarrow B}(C_A | e) \quad (4.64)$$

$$= \mathbb{K} \text{swap}_{A \leftrightarrow B}(C_A | e_B, \text{swap}_{B \leftrightarrow A}(x)_B^C) \quad \text{by locality} \quad (4.65)$$

$$= \mathbb{K} \text{swap}_{A \leftrightarrow B}(C_A | \text{swap}_{B \leftrightarrow A}(x)) \quad (4.66)$$

$$= \text{swap}_{B \leftrightarrow A} \mathbb{K} \text{swap}_{A \leftrightarrow B}(C_A | x) \quad (4.67)$$

$$= \mathbb{K}(C_A | x) \quad \text{by commutativity of exchange} \quad (4.68)$$

Because this holds for all  $x$ ,  $A \subset \mathbb{N}$ , by Lemma 4.2.5

$$\mathbb{Q} = \mathbb{K} \quad (4.69)$$

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

$$\mathbb{K}(A | e) = \text{swap}_c \mathbb{K} \text{swap}_c(A | e) \quad (4.70)$$

$$= \mathbb{K} \text{swap}_c(A | \text{swap}_c(e)) \quad (4.71)$$

$$= \mathbb{K} \text{swap}_c(A | e) \quad (4.72)$$

If: Suppose

$$\mathbb{K} = \begin{array}{c} \begin{array}{c} \triangleleft \\ \mu \end{array} \\ \begin{array}{c} X \text{ --- } \boxed{\mathbb{F}_{\text{ev}}} \text{ --- } Y \end{array} \end{array} \quad (4.73)$$

where  $\mu$  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  $A_S = \text{swap}_{[n] \leftrightarrow S} \times_{i \in [n]} A_i \times Y^{\mathbb{N}}$ ,  $A_T = \text{swap}_{S \leftrightarrow T}(A_S)$ ,  $B = (x_i i)_{i \in S}$  and  $C = (x_i i)_{i \in T} = (x_{\text{swap}_{S \leftrightarrow T}}(i) i)_{i \in S}$ . By Equations 4.58 and 4.62

$$\mathbb{K}(A_S | x) = \mu \text{swap}_{S \leftrightarrow B}(A_S) \quad (4.74)$$

$$= \mu \text{swap}_{T \leftrightarrow C}(A_T) \quad \text{by column exchangeability of } \mu \quad (4.75)$$

$$= \mathbb{K}(A_T | \text{swap}_{S \leftrightarrow T}(x)) \quad (4.76)$$

$$= \text{swap}_{S \leftrightarrow T} \mathbb{K}(A_T | x) \quad (4.77)$$

$$= \text{swap}_{S \leftrightarrow T} \mathbb{K} \text{swap}_{S \leftrightarrow T}(A_S | x) \quad (4.78)$$

so  $\mathbb{K}$  is causally contractible by Theorem 4.2.7.  $\square$

Theorem 4.2.11 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 4.2.11 is applied to probability set models.

**Definition 4.2.10** (Measurable set of probability distributions). Given a measurable set  $(\Omega, \mathcal{F})$ , the measurable set of distributions on  $\Omega$ ,  $\mathcal{M}_1(\Omega)$ , is the set of all probability distributions on  $\Omega$  equipped with the coarsest  $\sigma$ -algebra such that the evaluation maps  $\eta_B : \nu \mapsto \nu(B)$  are measurable for all  $B \in \mathcal{F}$ .

**Theorem 4.2.11.** *Given a kernel  $\mathbb{K} : X^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$ , let  $H := \mathcal{M}_1(Y^X)$  be the measurable set of probability distributions on  $(Y^X, \mathcal{Y}^X)$ .  $\mathbb{K}$  is causally contractible if and only if there is some  $\mathbb{H} : Y^{X \times \mathbb{N}} \rightarrow H$  and  $\mathbb{L} : H \times X \rightarrow Y$  such that*

$$\mathbb{K} = \begin{array}{c} \begin{array}{c} \triangleleft \\ \nu \end{array} \text{ --- } \boxed{\mathbb{F}_H} \text{ --- } \begin{array}{c} \boxed{\mathbb{L}} \text{ --- } Y \\ \begin{array}{c} X \text{ --- } \boxed{\mathbb{L}} \\ i \in \mathbb{N} \end{array} \end{array} \quad (4.79)$$

$$\iff \quad (4.80)$$

$$\mathbb{K}(\times_{i \in \mathbb{N}} A_i | (x_i)_{i \in \mathbb{N}}) = \int_H \prod_{i \in \mathbb{N}} \mathbb{L}(A_i | h, x_i) \mu \mathbb{F}_H(dh) \quad (4.81)$$

*Proof.* By Theorem 4.2.9, we can represent the conditional probability  $\mathbb{K}$  as

$$\mathbb{K} = \begin{array}{c} \triangleleft \mu \\ \text{---} X \text{---} \boxed{\mathbb{F}_{\text{ev}}} \text{---} Y \end{array} \quad (4.82)$$

where  $\mu$  is column exchangeable.

As a preliminary, we will show

$$\mathbb{F}_{\text{ev}} = \begin{array}{c} \boxed{\begin{array}{c} Y^D \text{---} \\ X \text{---} \end{array}} \boxed{\mathbb{F}_{\text{evs}}} \text{---} Y \\ i \in \mathbb{N} \end{array} \quad (4.83)$$

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

$$(x, (y_i)_{i \in X}) \mapsto y_x \quad (4.84)$$

Recall that  $\text{ev}$  is the function

$$((x_i)_{i \in \mathbb{N}}, (y_{ji})_{j \in X \times \mathbb{N}}) \mapsto (y_{xi})_{i \in \mathbb{N}} \quad (4.85)$$

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

$$\mathbb{F}_{\text{ev}}\left(\bigotimes_{i \in \mathbb{N}} A_i \mid (x_i)_{i \in \mathbb{N}}, (y_{ji})_{j \in X \times \mathbb{N}}\right) = \delta_{(y_{xi})_{i \in \mathbb{N}}} \left(\bigotimes_{i \in \mathbb{N}} A_i\right) \quad (4.86)$$

$$= \prod_{i \in \mathbb{N}} \delta_{y_{xi}}(A_i) \quad (4.87)$$

$$= \prod_{i \in \mathbb{N}} \mathbb{F}_{\text{evs}}(A_i \mid x_i, (y_{ji})_{j \in X}) \quad (4.88)$$

$$= \left( \bigotimes_{i \in \mathbb{N}} \mathbb{F}_{\text{evs}} \right) \left( \bigotimes_{i \in \mathbb{N}} A_i \mid (x_i)_{i \in \mathbb{N}}, (y_{ji})_{j \in X \times \mathbb{N}} \right) \quad (4.89)$$

which is what we wanted to show.

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

$$\mu = \begin{array}{c} \triangleleft \mu \\ \text{---} \boxed{\mathbb{F}_{\mathbb{H}}} \text{---} \boxed{\begin{array}{c} \text{---} \boxed{\mathbb{M}} \text{---} Y^D \\ i \in \mathbb{N} \end{array}} \end{array} \quad (4.90)$$

$$\iff \quad (4.91)$$

$$\mu\left(\bigotimes_{i \in \mathbb{N}} A_i\right) = \int_H \prod_{i \in \mathbb{N}} \mathbb{M}(A_i|h) \mu_{\mathbb{F}_{\mathbb{H}}}(\mathrm{d}h) \quad \forall A_i \in \mathcal{Y}^X \quad (4.92)$$

By Equations 4.82 and 4.83

$$\mathbb{K} = \begin{array}{c} \begin{array}{c} \nu \end{array} \begin{array}{c} \text{F}_H \end{array} \begin{array}{c} \boxed{\text{M}} \end{array} \begin{array}{c} \text{F}_{\text{ev}} \end{array} \begin{array}{c} Y \end{array} \\ \begin{array}{c} X \end{array} \begin{array}{c} i \in \mathbb{N} \end{array} \end{array} \quad (4.93)$$

$$:= \begin{array}{c} \nu \end{array} \begin{array}{c} \text{F}_H \end{array} \begin{array}{c} \boxed{\text{L}} \end{array} \begin{array}{c} Y \end{array} \\ \begin{array}{c} X \end{array} \begin{array}{c} i \in \mathbb{N} \end{array} \end{array} \quad (4.94)$$

Where we can connect the copied outputs of  $\mu\text{F}_H$  to the inputs of each  $\text{M}$  “inside the plate” as the plates in Equations 4.83 and 4.90 are equal in number and each connected wire represents a single copy of  $Y^D$ .

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

$$\mathbb{K}(\bigotimes_{i \in \mathbb{N}} A_i | x) = \int_H \prod_{i \in \mathbb{N}} \mathbb{L}(A_i | h, x_i) \mu(dh) \quad (4.95)$$

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  $A_T := \times_{i \in \mathbb{N}} B_{\text{swap}_{S \leftrightarrow T}(i)}$ .

$$\mathbb{K}(A_S | x) = \int_H \prod_{i \in S} \mathbb{L}(A_i | h, x_i) \mu(dh) \quad (4.96)$$

$$= \int_H \prod_{i \in T} \mathbb{L}(A_i | h, x_{\text{swap}_{S \leftrightarrow T}(i)}) \mu(dh) \quad (4.97)$$

$$= \text{swap}_{S \leftrightarrow T} \mathbb{K}(A_T | x) \quad (4.98)$$

$$= \text{swap}_{S \leftrightarrow T} \mathbb{K} \text{swap}_{S \leftrightarrow T}(A_S | x) \quad (4.99)$$

So by Theorem 4.2.7,  $\mathbb{K}$  is causally contractible.  $\square$

### 4.2.2 Causal contractibility with data-independent actions

Given a sequence of variables  $(D_i, Y_i)_{i \in \mathbb{N}}$  where the “inputs” are  $D := (D_i)_{i \in \mathbb{N}}$  and the “outputs” are  $Y = (Y_i)_{i \in \mathbb{N}}$ , say the inputs are independent of previous observations if  $D_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{<i}, C) | D_{<i}$  for all  $i \in \mathbb{N}$ . This models an experiment where it may be possible to choose different inputs  $D$ , but all the inputs are determined before the outputs  $Y$  are known. If a model satisfies this, then the dependence of  $Y$  on  $D$  is a sequence of repeatable response functions if and only if the uniform conditional  $\mathbb{P}_C^{Y|D}$  exists (Definition 2.4.3) and is causally contractible.

We call a model  $\mathbb{P}_C$  with sequential outputs  $Y$  and a corresponding sequence of data-independent inputs  $D$  a “sequential just-do model”.



**Definition 4.2.12** (Sequential just-do model). A *sequential just-do model* is a triple  $(\mathbb{P}_C, D, Y)$  where  $\mathbb{P}_C$  is a probability set on  $(\Omega, \mathcal{F})$ ,  $D$  is a sequence of “inputs”  $D := (D_i)_{i \in \mathbb{N}}$  and  $Y$  is a corresponding sequence of “outputs”  $Y = (Y_i)_{i \in \mathbb{N}}$  where  $D_i : \Omega \rightarrow D$  and  $Y_i : \Omega \rightarrow Y$ . Furthermore, it is required that  $X_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{<i}, C) | X_{<i}$  for all  $i \in \mathbb{N}$ , and  $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e C | D$ .

To apply Theorem 4.2.11 to a sequential just-do model, it is necessary to extend the model to a larger sample space including “latent variables” taking values in  $Y^D$ . A latent extension is a model over a larger collection of variables that reduces to the original model when we restrict our attention to the original collection of variables.

**Definition 4.2.13** (Latent extension). Given a probability set  $\mathbb{P}'_C$  on  $(\Omega, \mathcal{F})$  and some measurable set  $(G, \mathcal{G})$ , a probability set  $\mathbb{P}_C$  is a *latent extension* of  $\mathbb{P}'_C$  to  $(\Omega \times G, \mathcal{F} \otimes \mathcal{G})$  if  $\mathbb{P}_C \mathbb{F}_{\Pi_\Omega} = \mathbb{P}'_C$ .

**Nxample 4.2.14** (Variables on a latent extension). Given a probability set  $\mathbb{P}'_C$  on  $(\Omega, \mathcal{F})$  and a latent extension  $\mathbb{P}_C$  on  $(\Omega \times G, \mathcal{F} \otimes \mathcal{G})$ , every variable on the original sample space is given a primed name  $X', Y'$  etc., and corresponds to an unprimed variable on the larger space  $X := \Pi_\Omega \circ X'$ .

Theorem 4.2.15 applies Theorem 4.2.11 to the case of a model with data-independent actions and derives the required conditional independences and equalities to show that a sequential just-do model  $(\mathbb{P}_C, D, Y)$  with causally contractible  $\mathbb{P}_C^{Y|X}$  satisfies the required conditional independences and equalities of conditional distributions for  $X$  and  $Y$  to be related by repeatable response functions.

**Theorem 4.2.15** (Data-independent causal contractibility). *Given a sequential just-do model  $(\mathbb{P}'_C, D', Y')$  on  $(\Omega, \mathcal{F})$ , then  $\mathbb{P}'_C^{Y'|D'}$  is causally contractible if and only if there is a latent extension  $\mathbb{P}_C$  of  $\mathbb{P}'_C$  to  $(\Omega \times Y^{D \times \mathbb{N}}, \mathcal{F} \otimes \mathcal{Y}^{D \times \mathbb{N}})$  with some hypothesis  $H : \Omega \times Y^{D \times \mathbb{N}} \rightarrow H$  such that  $Y_i \perp\!\!\!\perp_{\mathbb{P}'_C}^e (Y_{<i}, X_{<i}, C) | (X_i, H)$  and  $\mathbb{P}_C^{Y_i | X_i H} = \mathbb{P}_C^{Y_j | X_j H}$  for all  $i, j \in \mathbb{N}$  and  $H \perp\!\!\!\perp_{\mathbb{P}_C} (X, C)$ .*

*Proof.* If: First, define the extension  $\mathbb{P}_C$ . From Theorem 4.2.9 and causal contractibility of  $\mathbb{P}'_C^{Y'|D'}$  there is some  $\mu \in \Delta(Y^{D \times \mathbb{N}})$  such that

$$\mathbb{P}'_C^{Y'|D'} = \begin{array}{c} \triangle \mu \\ \swarrow \quad \searrow \\ D' \quad \text{---} \quad \boxed{\mathbb{F}_{\text{ev}}} \quad \text{---} \quad Y' \end{array} \quad (4.100)$$

Let  $\mathbb{P}_C^{Y^D | D} = \mu \otimes \text{del}_{D^{\mathbb{N}}}$ ,  $\mathbb{P}_C^{Y | Y^D} = \mathbb{F}_{\text{ev}}$  and  $Y^D := \Pi_{Y^{D \times \mathbb{N}}}$ , the projection  $\Omega \times Y^{D \times \mathbb{N}} \rightarrow \Omega$ . Let  $W = \Pi_\Omega$  and for each  $\alpha \in C$ , set

$$\mathbb{P}_\alpha^{W | Y^D} = \begin{array}{c} D \quad \text{---} \quad \bullet \\ \swarrow \quad \searrow \\ Y^D \quad \text{---} \quad \boxed{\mathbb{P}_C^{Y | DY^D}} \quad \text{---} \quad \boxed{\mathbb{P}_\alpha^{W | DY}} \quad \text{---} \quad W \end{array} \quad (4.101)$$

Then

$$\mathbb{P}_\alpha^W = \begin{array}{c} \triangleleft \mathbb{P}_\alpha^{D'} \\ \triangleleft \mathbb{P}_C^{Y^D} \end{array} \begin{array}{c} \bullet \\ \curvearrowright \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y|DY^D}} \\ \boxed{\mathbb{P}_\alpha^{W|DY}} \end{array} \begin{array}{c} \text{---} W \end{array} \quad (4.102)$$

$$= \begin{array}{c} \triangleleft \mathbb{P}_\alpha^{D'} \end{array} \begin{array}{c} \bullet \\ \curvearrowright \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y'|D'}} \\ \boxed{\mathbb{P}_\alpha^{W'|D'Y'}} \end{array} \begin{array}{c} \text{---} W \end{array} \quad (4.103)$$

$$= \mathbb{P}_\alpha^{\text{id}_\Omega} = \mathbb{P}_\alpha \quad (4.104)$$

Before going further, it's necessary to check that there is some nonempty probability set  $\mathbb{P}_C$  with these conditionals. By Theorem 2.4.24, because  $\mathbb{P}_\alpha^W = \mathbb{P}'_\alpha$  it is sufficient to show that  $\mathbb{P}_\alpha^{Y^D|W}$  is valid. Because

$$(W, Y^D)(\Omega \times Y^{D \times \mathbb{N}}) = \Omega \times Y^{D \times \mathbb{N}} \quad (4.105)$$

$$= W(\Omega \times Y^{D \times \mathbb{N}}) \times Y^D(\Omega \times Y^{D \times \mathbb{N}}) \quad (4.106)$$

there are no impossible events, and so validity is guaranteed for any  $\mathbb{P}_\alpha^{Y^D|W}$ .

Thus  $\mathbb{P}_C$  is a latent extension of  $\mathbb{P}'_C$ , and so  $\mathbb{P}_C^{Y|D}$  is also causally contractible.

From Theorem 4.2.11 and by construction of  $\mathbb{P}_C$ , there exists a directing random measure  $H^* : Y^{D \times \mathbb{N}} \rightarrow H$  such that, defining  $H = H^* \circ \Pi_{D \times \mathbb{N}}$

$$\mathbb{P}_C^{Y|D} = \begin{array}{c} \triangleleft \mathbb{P}_C^H \end{array} \begin{array}{c} \bullet \\ \curvearrowright \end{array} \begin{array}{c} \boxed{\mathbb{L}} \\ \text{---} Y_i \end{array} \begin{array}{c} D_i \\ i \in \mathbb{N} \end{array} \quad (4.107)$$

it remains to be shown that  $\mathbb{L}$  is a version of  $\mathbb{P}^{Y_i|D_i,H}$  for all  $i \in \mathbb{N}$  and  $Y_i \perp\!\!\!\perp_{\mathbb{P}'_C}^e (Y_{<i}, D_{<i}, C) | (D_i, H)$ .

To show  $\mathbb{L}$  is a version of  $\mathbb{P}^{Y_i|HD_i}$  for all  $i \in \mathbb{N}$ :

$$\mathbb{L} = \begin{array}{c} \begin{array}{c} H \\ D_i \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y^D|H}} \\ \boxed{\mathbb{F}_{\text{evs}}} \end{array} \begin{array}{c} \\ \end{array} \begin{array}{c} \\ Y_i \end{array} \end{array} \quad (4.108)$$

$$= \begin{array}{c} \begin{array}{c} H \\ D_i \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y_i^D|H}} \\ \boxed{\mathbb{P}_C^{Y_i|Y_i^D D_i}} \end{array} \begin{array}{c} \\ \end{array} \begin{array}{c} \\ Y_i \end{array} \end{array} \quad (4.109)$$

$$= \begin{array}{c} H \\ D_i \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y_i^D|HD_i}} \\ \boxed{\mathbb{P}_C^{Y_i|Y_i^D D_i}} \end{array} \begin{array}{c} \\ \end{array} \begin{array}{c} \\ Y_i \end{array} \quad (4.110)$$

$$= \begin{array}{c} H \\ D_i \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y_i^D|HD_i}} \\ \boxed{\mathbb{P}_C^{Y_i|Y_i^D D_i H}} \end{array} \begin{array}{c} \\ \end{array} \begin{array}{c} \\ Y_i \end{array} \quad (4.111)$$

$$= \mathbb{P}_C^{Y_i|HD_i} \quad (4.112)$$

Where 4.110 follows from  $H \perp\!\!\!\perp_{\mathbb{P}_C}^e (D_i, C)$ , which itself follows from  $Y_i^D \perp\!\!\!\perp_{\mathbb{P}_C}^e (D_i, C)$  which holds by construction. 4.111 follows from  $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (H, C) | (Y_i^D, D_i)$ , which follows from  $Y_i$  being a deterministic function of  $(Y_i^D, D_i)$ .

For independence, note that

$$\mathbb{P}_C^{Y_{< i} | HX_{< i} X_i} = \begin{array}{c} \begin{array}{c} X_{< i}^H \\ X_i \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y_{< i} | HX_{< i}}} \\ \boxed{\mathbb{P}_C^{Y_{< i} | HX_{< i}}} \end{array} \begin{array}{c} \\ * \end{array} \begin{array}{c} \\ Y_{< i} \end{array} \end{array} \quad (4.113)$$

$$= \begin{array}{c} \begin{array}{c} X_{< i}^H \\ X_i \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y_{< i} | HX_{< i}}} \\ * \end{array} \begin{array}{c} \\ Y_{< i} \end{array} \end{array} \quad (4.114)$$

hence  $Y_{< i} \perp\!\!\!\perp_{\mathbb{P}_C}^e (X_i, C) | (H, X_{< i})$

Then

$$\mathbb{P}_C^{Y_i Y_{< i} | HD_i D_{< i}} = \begin{array}{c} \begin{array}{c} H \\ X_{< i} \\ X_i \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y_{< i} | HX_{< i}}} \\ * \end{array} \begin{array}{c} \\ * \end{array} \begin{array}{c} \\ Y_{< i} \end{array} \end{array} \quad (4.115)$$

$$\Rightarrow \mathbb{P}_C^{Y_i | HD_i D_{< i}} \cong \begin{array}{c} \begin{array}{c} H \\ X_i \end{array} \begin{array}{c} \boxed{\mathbb{P}_C^{Y_i | HX_i}} \\ * \end{array} \begin{array}{c} \\ Y_i \end{array} \end{array} \quad (4.116)$$

by Theorem 2.4.32. Hence  $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (X_{< i}, Y_{< i}, C) | (H, X_i)$ .

Only if: If  $\mathbb{P}_C$  is a latent extension of  $\mathbb{P}'_C$ , then  $\mathbb{P}_C^{Y|D}$  is causally contractible if and only if  $\mathbb{P}'_C^{Y'|D'}$  is causally contractible. Thus it is sufficient to show  $\mathbb{P}_C^{Y|D}$  is causally contractible.

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

$$\mathbb{P}_C^{Y_i | HX_{[i]} Y_{<i}} \stackrel{\mathbb{P}_C}{\cong} \text{del}_{X^{i-1} \times Y^{i-1}} \otimes \mathbb{P}_C^{Y_1 | HX_1} \quad (4.117)$$

Thus for all  $n \in \mathbb{N}$  by repeated application of Theorem 2.4.32

$$\mathbb{P}_C^{Y_{[n]} | HX_{[n]}} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} \text{H} \text{---} \bullet \text{---} \boxed{\mathbb{L}} \text{---} Y_i \\ \text{D}_i \text{---} \curvearrowright \\ i \in [n] \end{array} \quad (4.118)$$

thus by Lemma 4.2.5

$$\mathbb{P}_C^{Y_{\mathbb{N}} | HX_{\mathbb{N}}} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} \text{H} \text{---} \bullet \text{---} \boxed{\mathbb{L}} \text{---} Y_i \\ \text{D}_i \text{---} \curvearrowright \\ i \in \mathbb{N} \end{array} \quad (4.119)$$

and, because  $H \perp_{\mathbb{P}_C}^e (X, C)$

$$\mathbb{P}_C^{Y_{\mathbb{N}} | X_{\mathbb{N}}} \stackrel{\mathbb{P}_C}{\cong} \begin{array}{c} \triangle \mathbb{P}_C^H \text{---} \bullet \text{---} \boxed{\mathbb{L}} \text{---} Y_i \\ \text{D}_i \text{---} \curvearrowright \\ i \in \mathbb{N} \end{array} \quad (4.120)$$

causal contractibility follows from Theorem 4.2.11.  $\square$

A consequence of Theorem 4.2.7 applied to a just-do models  $\mathbb{P}_C$  with causally contractible  $\mathbb{P}_C^{Y|D}$  is that, for any  $A, B \subset \mathbb{N}$  with  $|A| = |B|$ ,  $\mathbb{P}_C^{Y_A | D_A} = \mathbb{P}_C^{Y_B | D_B}$ . A further consequence is the interchangeability of conditioning data – for any  $i \in \mathbb{N}$ ,  $\mathbb{P}_C^{Y_i | D_i Y_A D_A} = \mathbb{P}_C^{Y_i | D_i Y_B D_B}$ .

**Theorem 4.2.16** (Equality of subsequence conditionals). *A sequential just-do model  $(\mathbb{P}_C, D, Y)$  with  $\mathbb{P}_C^{Y|D}$  causally contractible satisfies, for any  $A, B \subset \mathbb{N}$  with  $|A| = |B|$*

$$\mathbb{P}_C^{Y_A | D_A} \stackrel{\mathbb{P}_C}{\cong} \mathbb{P}_C^{Y_B | D_B} \quad (4.121)$$

*Proof.* Only if: For any  $A, B \subset \mathbb{N}$ , let  $\text{swap}_{B \leftrightarrow A, D} : D^{\mathbb{N}} \rightarrow D^{\mathbb{N}}$  be the transposition of  $B$  with  $A$  indices and  $\text{swap}_{B \leftrightarrow A, Y} : Y^{\mathbb{N}} \rightarrow Y^{\mathbb{N}}$  be the same defined on  $Y$ . By Theorem 4.2.7

$$\mathbb{P}_C^{Y_A | D_A} \otimes \text{del}_{D^{\mathbb{N}}} = \mathbb{P}_C^{Y|D} \text{marg}_A \quad (4.122)$$

$$= \text{swap}_{A \leftrightarrow [n], D} \mathbb{P}_C^{Y_{[n]} | D} \quad (4.123)$$

$$= \mathbb{P}_C^{Y_{[n]} | D_{[n]}} \otimes \text{del}_{D^{\mathbb{N}}} \quad (4.124)$$

$$= \text{swap}_{N \leftrightarrow [n], D} \mathbb{P}_C^{Y_{[n]} | D} \quad (4.125)$$

$$= \mathbb{P}_C^{Y_B | D_B} \otimes \text{del}_{D^{\mathbb{N}}} \quad (4.126)$$

$\square$

### 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  $(D_i, Y_i)_{i \in \mathbb{N}}$  has  $\mathbb{P}_C^{Y|D}$  causally contractible. This follows from the fact that

$$\mathbb{P}_C^{YD} = \begin{array}{c} \begin{array}{c} \triangleleft \mathbb{P}_C^H \end{array} \quad \begin{array}{|c|} \hline \begin{array}{c} \begin{array}{c} \bullet \end{array} \begin{array}{c} \begin{array}{c} \mathbb{P}_C^{Y_i|D_i} \end{array} \end{array} \end{array} \begin{array}{c} \begin{array}{c} Y_i \\ D_i \end{array} \end{array} \\ \hline \end{array} \quad i \in \mathbb{N} \end{array} \quad (4.127)$$

and so

$$\begin{array}{c} \triangleleft \mathbb{P}_C^H \end{array} \quad \begin{array}{|c|} \hline \begin{array}{c} \begin{array}{c} \bullet \end{array} \begin{array}{c} \mathbb{P}_C^{Y_i|D_i} \end{array} \end{array} \begin{array}{c} Y_i \\ D_i \end{array} \\ \hline \end{array} \quad i \in \mathbb{N} \quad (4.128)$$

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

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  $(D, Y)_{i \in \mathbb{N}}$  and the active interventions are  $(E, Z)_{i \in \mathbb{N}}$ . By the previous argument,  $\mathbb{P}_C^{Y|D}$  is causally contractible. We might further assume that  $\mathbb{P}_C^{YZ|DE}$  is causally contractible – that is, there is a repeatable response function  $\mathbb{P}_C^{Z_i|E_i H}$  equal to  $\mathbb{P}_C^{Y_i|D_i H}$ .

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

$$\mathbb{P}_\alpha^{DE \text{ swap}_\rho} = \mathbb{P}_\alpha^{DE} \quad (4.129)$$

then by commutativity of exchange

$$\mathbb{P}_\alpha^{YZ} = \mathbb{P}_\alpha^{DE} \mathbb{P}_C^{YZ|DE} \quad (4.130)$$

$$= \mathbb{P}_\alpha^{DE \text{ swap}_\rho} \mathbb{P}_C^{YZ|DE} \quad (4.131)$$

$$= \mathbb{P}_\alpha^{DE} \mathbb{P}_C^{YZ|DE \text{ swap}_\rho} \quad (4.132)$$

$$= \mathbb{P}_C^{YZ|DE \text{ swap}_\rho} \quad (4.133)$$

However, the assumption that  $\mathbb{P}_C^{YZ|DE}$  is causally contractible seems unreasonable in most situations. One implication of this assumption is (by Theorem 4.2.7):

$$\mathbb{P}_C^{YZ_i|DE_i} = \mathbb{P}^{Z|E} \quad (4.134)$$

$$\implies \mathbb{P}_C^{Z_i|E_i DY} = \mathbb{P}^{Z_i|E_i E_{\{i\}^C} Z_{\{i\}^C}} \quad (4.135)$$

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

- $X_i \perp\!\!\!\perp_{\mathbb{P}_C}^e D_i C | H$  ( $X_i$  is extended independent of  $D_i$  conditional on  $H$ )
- $\mathbb{P}_C^{X_i|H} \cong \mathbb{P}_C^{X_1|H}$  (the distribution of  $X$  is exchangeable)

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

$$\mathbb{P}_\alpha^{Y_i|D_i H}(A|d, h) = \int_X \mathbb{P}_\alpha^{Y_i|X_i D_i H}(A|d, x, h) \mathbb{P}_\alpha^{X_i|D_i H}(dx|d, h) \quad (4.136)$$

$$= \int_X \mathbb{P}_C^{Y_i|X_i D_1 H}(A|d, x, h) \mathbb{P}_C^{X_i|H}(dx|h) \quad (4.137)$$

$$= \int_X \mathbb{P}_C^{Y_1|X_1 D_1 H}(A|d, x, h) \mathbb{P}_C^{X_1|H}(dx|h) \quad (4.138)$$

Equation 4.138 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.

### 4.3 Causal contractibility in sequences of active choices

Assessing when a particular sequence of experiments should be modeled with a causally contractible model can be difficult. As noted, a purely observational sequence is causally contractible if it is exchangeable. However, the point of all this theory is to study models that offer different choices. The assumption of causal contractibility can be justified for a sequence of active interventions if the following are satisfied:

1. There exist variables  $I$  representing “unique experiment identifiers” which satisfy the assumption that  $\mathbb{P}_C^{Y|DI}$  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. Given a permutation  $\rho$  of identifiers,  $\mathbb{P}_\alpha^{YD\rho(I)} = \mathbb{P}_\alpha^{YDI}$  (informally: unique identifiers are not themselves informative)
3. The map  $\alpha \rightarrow \mathbb{P}_\alpha^D$  is deterministic Markov kernel associated with an invertible function  $f : C \rightarrow D^{\mathbb{I}}$

### 4.3. CAUSAL CONTRACTIBILITY IN SEQUENCES OF ACTIVE CHOICES 79

Theorem 4.3.4 shows that, under these assumptions,  $\mathbb{P}_C^{Y|D}$  is also causally contractible.

The first assumption – that causal contractibility is satisfied jointly conditioning on decisions  $D$  and identifiers  $I$  – seems to often be a background assumption in the literature, while assumptions similar to the second two are discussed explicitly. For example, Greenland and Robins (1986) explain

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

Note that exchanging individuals involved in an experiment and exchanging the individuals’ exposure states are two different things, and the former doesn’t imply the latter – for example, there might be some background trend such that individuals treated later experience different outcomes to individuals treated at the start. Assumptions 1 and 2 *together* imply that permuting identifiers or permuting decisions both lead to the same distribution.

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.

This seems on the face of it similar to assumption 2: that I can permute the identifies “me” and “someone else” without changing the model.

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

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

Here “the science” means (roughly) *the response function of each individual*, and exchangeability of these response functions is a similar assumption to permutability of individual identifiers (though we don’t derive the exact correspondence here). Rubin’s second condition is that treatment assignment is ignorable. Like Assumption 3, this assumption limits “how much we can learn from the treatment assignment”, but again we don’t derive the exact correspondence. Note that in Rubin’s scheme the preliminary assumption of *stable unit-treatment values* is made in order to establish the existence of individual response functions, which plays a similar role to Assumption 1.

Rubin’s result also differs substantially from this one in that it applies to *identification of potential outcomes in randomised experiments*, while this result is about the existence of response conditionals *when there are different choices that can be made*.

As an example of the application of Theorem 4.3.4, consider an experiment where  $n$  patients, each with an individual identifier  $l_i$ , receive treatment  $D_i$  and experience outcome  $Y_i$ .  $\mathbb{P}_C^{Y_i|D_i l_i}$  can be extended to an infinite sequence  $\mathbb{P}_C^{Y|D l}$  that is causally contractible (see Assumption 1), no matter which choice  $\alpha \in C$  is decided on, 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  $\mathbb{P}_C^{Y|D}$  is causally contractible, and hence there are response functions  $\mathbb{P}_C^{Y_i|D_i H}$ .

Theorem 4.3.4 can also be extended to the case where  $D$  is a function of the choice  $\alpha$  and a “random signal”  $R$ .

**Lemma 4.3.1.** *Given sequential just-do model  $(\mathbb{P}_C, (D, l), Y)$  with  $\mathbb{P}_C^{Y|D l}$  causally contractible, if  $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e (l, C) | D$  then  $\mathbb{P}_C^{Y|D}$  is also causally contractible.*

*Proof.* For arbitrary  $\nu \in \Delta(I^{\mathbb{N}})$ , by assumption of causal contractibility of  $\mathbb{P}_C^{Y|D l}$  and Theorem 4.2.11

$$\mathbb{P}_C^{Y|D l} = \begin{array}{c} \triangle \mathbb{P}_C^H \\ \begin{array}{c} D \\ l \end{array} \rightarrow \begin{array}{c} \Pi_{D,i} \\ \Pi_{l,i} \end{array} \rightarrow \begin{array}{c} \mathbb{P}_C^{Y_i|HD_i l_i} \\ Y_i \end{array} \\ i \in \mathbb{N} \end{array} \quad (4.139)$$

$$= \begin{array}{c} \triangle \mathbb{P}_C^H \\ \begin{array}{c} D \\ l \end{array} \rightarrow \begin{array}{c} \Pi_{D,i} \\ \Pi_{l,i} \end{array} \rightarrow \begin{array}{c} \mathbb{P}_C^{Y_i|HD_i l_i} \\ Y_i \end{array} \\ l \rightarrow * \triangle \nu \\ i \in \mathbb{N} \end{array} \quad (4.140)$$

$$= \begin{array}{c} \triangle \mathbb{P}_C^H \\ \begin{array}{c} D \\ l \end{array} \rightarrow \begin{array}{c} \Pi_{D,i} \\ \Pi_{l,i} \end{array} \rightarrow \begin{array}{c} \mathbb{P}_C^{Y_i|HD_i l_i} \\ Y_i \end{array} \\ l \rightarrow * \triangle \nu_1 \\ i \in \mathbb{N} \end{array} \quad (4.141)$$

$$\Rightarrow \mathbb{P}_C^{Y|D} = \begin{array}{c} \triangle \mathbb{P}_C^H \\ D \rightarrow \begin{array}{c} \Pi_{D,i} \\ \Pi_{l,i} \end{array} \rightarrow \begin{array}{c} \mathbb{P}_C^{Y_i|HD_i l_i} \\ Y_i \end{array} \\ \triangle \nu_1 \\ i \in \mathbb{N} \end{array} \quad (4.142)$$

Applying Theorem 4.2.11,  $\mathbb{P}_C^{Y|D}$  is causally contractible.  $\square$

An *identifier variable* is a variable  $l$  that takes values in the set of finite permutations of  $\mathbb{N}$ . It is associated with a sequence  $(l_i)_{i \in \mathbb{N}}$  where  $l_i = l(i)$ . Each  $l_i$  takes values in  $\mathbb{N}$  and  $l_i \neq l_j$  for all  $j \neq i$ .



**Definition 4.3.2** (Identifier variable). Given a probability set  $\mathbb{P}_C$  on  $(\Omega, \mathcal{F})$ , let  $I$  be the set of finite permutations  $\mathbb{N} \rightarrow \mathbb{N}$ . A variable  $\mathbf{l} : \Omega \rightarrow I$  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 4.3.3.** *Given a probability set  $\mathbb{P}_C$  where  $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e C|(D, \mathbf{l})$  and  $\mathbf{l} : \Omega \rightarrow I$  is an identifier variable, if for each finite permutation  $\rho : \mathbb{N} \rightarrow \mathbb{N}$*

$$\mathbb{P}_\alpha^{Y|\mathbf{D}} = (\text{swap}_{\rho(I)} \otimes \text{Id}_X) \mathbb{P}_\alpha^{Y|\mathbf{D}} \quad (4.143)$$

then  $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e (\mathbf{l}, C)|D$ .

*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  $\rho \circ \rho^{-1} = \text{id}_\mathbb{N}$ . Then

$$\mathbb{P}_C^{Y|\mathbf{D}}(B|i, d) = (\mathbb{F}_{\rho^{-1}} \otimes \text{Id}_X) \mathbb{P}_C^{Y|\mathbf{D}}(B|\rho, d) \quad (4.144)$$

$$= \mathbb{P}_C^{Y|\mathbf{D}}(B|\text{id}_\mathbb{N}, d) \quad (4.145)$$

Therefore

$$\mathbb{P}_C^{Y|\mathbf{D}} \stackrel{\mathbb{P}_C}{\cong} \text{erase}_I \otimes \mathbb{K} \quad (4.146)$$

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

$$(B|d) \mapsto \mathbb{P}_C^{Y|\mathbf{D}}(B|\text{id}_\mathbb{N}, d) \quad (4.147)$$

□

The following theorem assumes that the set of choices  $C$  is countable and there is a one-to-one function  $f : C \rightarrow D^\mathbb{N}$ . Thus, if  $|D| > 1$ ,  $f$  cannot be surjective.

**Theorem 4.3.4.** *Given a sequential just-do model  $(\mathbb{P}_C, (D, \mathbf{l}), Y)$  on  $(\Omega, \mathcal{F})$  with  $C$  countable,  $\mathbb{P}_C^{Y|\mathbf{D}}$  causally contractible  $\mathbf{l} : \Omega \rightarrow I$  an identifier variable, if for each  $\alpha \in C$ ,  $\rho \in I$*

$$\mathbb{P}_\alpha^{Y|\mathbf{l}} = \mathbb{F}_\rho \mathbb{P}_\alpha^{Y|\mathbf{l}} \quad (4.148)$$

and furthermore

$$Y \perp\!\!\!\perp_{\mathbb{P}_C}^e C|D \quad (4.149)$$

$$Y \perp\!\!\!\perp_{\mathbb{P}_C}^e D|C \quad (4.150)$$

then  $\mathbb{P}_C^{Y|\mathbf{D}}$  is causally contractible.

$$= \text{Diagram: } I \text{ enters a triangle labeled } \mathbb{P}_\alpha^D \text{ from the left. The output of the triangle enters a rectangle labeled } \mathbb{P}_C^{Y|ID} \text{ from the left. The output of the rectangle is } Y. \quad (4.152)$$

$$Q^{Y|IC} \text{ is } \begin{array}{c} I \\ \text{---} \\ C \end{array} \begin{array}{c} \boxed{Q^{D|C}} \\ \text{---} \end{array} \boxed{P_C^{Y|ID}} \text{---} Y \quad (4.153)$$

By assumption  $\forall I \perp_{\mathbb{P}_C}^e D|C$ , it is also the case that

$$\begin{array}{c} I \\ D \end{array} \begin{array}{c} \text{---} \\ \text{---} \end{array} \begin{array}{c} \boxed{Q^Y|C} \\ \boxed{Q^C|D} \end{array} \text{---} Y \quad (4.155)$$

$$\text{I} \xrightarrow{\quad} \boxed{\mathbb{P}_C^{Y|D}} \xrightarrow{\quad} Y$$

(4.156)

But

$$= \mathbb{P}_C^{Y|ID} \Rightarrow \begin{array}{c} I \text{ ---} \\ D \text{ ---} \end{array} \begin{array}{|c|c|} \hline Q^{C|D} & Q^{D|C} \\ \hline \end{array} \begin{array}{|c|} \hline \mathbb{P}_C^{Y|ID} \\ \hline \end{array} \text{ --- } Y = \mathbb{P}_C^{Y|ID} \quad (4.158)$$

Furthermore, by assumption, for any permutation  $\rho : \mathbb{N} \rightarrow \mathbb{N}$

$$\mathbb{Q}^{Y|IC} = \begin{array}{c} I \text{ --- } \boxed{\mathbb{F}_\rho} \\ D \text{ --- } \end{array} \boxed{\mathbb{Q}^{Y|IC}} \text{ --- } Y \quad (4.159)$$

$$\Rightarrow \mathbb{P}_C^{Y|D} = \begin{array}{c} I \text{ --- } \boxed{\mathbb{F}_\rho} \\ D \text{ --- } \boxed{\mathbb{Q}^{C|D}} \text{ --- } \boxed{\mathbb{Q}^{D|C}} \end{array} \boxed{\mathbb{P}_C^{Y|D}} \text{ --- } Y \quad (4.160)$$

$$= \begin{array}{c} I \text{ --- } \boxed{\mathbb{F}_\rho} \\ D \text{ --- } \boxed{\mathbb{Q}^{C|D}} \text{ --- } \boxed{\mathbb{Q}^{D|C}} \end{array} \boxed{\mathbb{P}_C^{Y|D}} \text{ --- } Y \quad (4.161)$$

$$= \begin{array}{c} I \text{ --- } \boxed{\mathbb{F}_\rho} \\ D \text{ --- } \end{array} \boxed{\mathbb{P}_C^{Y|IC}} \text{ --- } Y \quad (4.162)$$

Then by Lemma 4.3.3 the independence  $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e IC|D$  holds, and by Lemma 4.3.1  $\mathbb{P}_C^{Y|D}$  is causally contractible.  $\square$

Theorem 4.3.4 can be extended to the case where decisions  $D$  are one-to-one deterministic, or random mixtures of one-to-one deterministic.

**Corollary 4.3.5.** *Given a sequential just-do model  $(\mathbb{P}_{C'}, (D, I), Y)$  satisfying the conditions of Theorem 4.3.4 and a second model  $(\mathbb{P}_C, (D, I), Y)$  such that for all  $\alpha \in C$  there is some set of coefficients  $k_i$  such that*

$$\mathbb{P}_\alpha = \sum_{c \in C'} k_c \mathbb{P}_c \quad (4.163)$$

then  $\mathbb{P}_C^{Y|D}$  is also causally contractible.

*Proof.* For all  $\alpha \in C$

$$\mathbb{P}_\alpha^{Y|D} = \sum_{c \in C'} k_c \mathbb{P}_c^{Y|D} \quad (4.164)$$

$$= \mathbb{P}_{C'}^{Y|D} \quad (4.165)$$

$\square$

Dropping the assumption  $YI \perp\!\!\!\perp_{\mathbb{P}_C}^e C|D$  means that, in general, one or both of  $\mathbb{P}_C^{Y|D}$  or  $\mathbb{P}_C^{Y|ID}$  may be ill-defined (note that the independence is merely a sufficient condition, not a necessary condition for these uniform conditional probabilities). The condition  $YI \perp\!\!\!\perp_{\mathbb{P}_C}^e C|D$  alone also does *not* imply the conclusion of Theorem 4.3.4.

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  $D \rightarrow Y$ . This generates space of hypotheses  $H = \{0, 1\}^{\mathbb{N}}$ , which needs to be equipped

with an algebra of measurable sets. 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 4.3.6.** Take  $Y = C = D = \{0, 1\}$  and take  $(H, \mathcal{H})$  to be  $\{0, 1\}^{\mathbb{N}}$  equipped with the product topology. For any  $i \neq 1$ ,  $Y_i \perp_i D_i \perp_{\mathbb{P}_C} C$ , while  $\mathbb{P}_\alpha^{D_1} = \delta_\alpha$  and  $I_i \perp_{\mathbb{P}_C} C$ .

$Y \perp_{\mathbb{P}_C} C | D$  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

$$\mathbb{P}_C^{Y_i | H_i D_i}(y|h, j, d) = \delta_1(p(j, h))\delta_d(y) + \delta_0(p(j, h))\delta_{1-d}(y) \quad (4.166)$$

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

$$Y_i \perp_{\mathbb{P}_C} (X_{<i}, Y_{<i}, I_{<i}, C) | (X_i, Y_i, H) \quad (4.167)$$

Then  $\mathbb{P}_C^{Y | D^I}$  is causally contractible. Set  $\mathbb{P}_C^H$  to be the uniform measure on  $(H, \mathcal{H})$  and for  $i > 1$

$$\mathbb{P}_C^{D_i | I_i H}(d|j, h) = \delta_{p(j, h)}(d) \quad (4.168)$$

that is, if  $h$  maps  $j$  to 1,  $D$  is 1 while if  $h$  maps  $j$  to 0,  $D$  is 0. This also implies

$$\mathbb{P}_C^{I_i | D_i H}(p(\cdot, h)^{-1}(d)|d, h) = 1 \quad (4.169)$$

Then, for  $i > 1$

$$\mathbb{P}_\alpha^{Y_i | H D_i}(y|h, d) = \sum_{j \in \mathbb{N}} \delta_1(p(j, h))\delta_d(y)\mathbb{P}_C^{I_i | D_i H}(j|d, h) + \delta_0(p(j, h))\delta_{1-d}(y)\mathbb{P}_C^{I_i | D_i H}(j|d, h) \quad (4.170)$$

$$= \sum_{j \in \mathbb{N}} \delta_1(d)\delta_d(y)\mathbb{P}_C^{I_i | D_i H}(j|d, h) + \delta_0(d)\delta_{1-d}(y)\mathbb{P}_C^{I_i | D_i H}(j|d, h) \quad \text{by Eq 4.169} \quad (4.171)$$

$$= \delta_1(y) \quad (4.172)$$

$$\implies \mathbb{P}_\alpha^{Y_i | D_i}(y|d) = \delta_1(y) \quad (4.173)$$

For  $q \in I$ , set

$$\mathbb{P}_C^{I | H}(q|h) = \begin{cases} 0.5 & q = (1, 2, 3, 4, \dots) \text{ or } (1, 3, 2, 4, \dots) \\ 0 & \text{otherwise} \end{cases} \quad (4.174)$$

and set

$$\mathbb{P}_C^{H | D}(h) = \begin{cases} 0.5 & h = (0, 1, 0, 1, 1, \dots) \text{ or } h = (0, 0, 1, 1, 1, \dots) \\ 0 & \text{otherwise} \end{cases} \quad (4.175)$$

Let  $\overline{H}$  be the support of  $\mathbb{P}_C^{\mathbf{H}|\mathbf{D}}(h)$ .

Then for  $i = 1$

$$\mathbb{P}_\alpha^{\mathbf{Y}_1|\mathbf{D}_1}(y|h, d) = \sum_{h \in \overline{H}} \sum_{j \in \mathbb{N}} \mathbb{P}_\alpha^{\mathbf{I}_1|\mathbf{D}_1\mathbf{H}}(j|d, h) \mathbb{P}_C^{\mathbf{H}|\mathbf{D}_1}(h|d) (\delta_1(p(j, h))\delta_d(y) + \delta_0(p(j, h))\delta_{1-d}(y)) \quad (4.176)$$

$$= \sum_{h \in \overline{H}} 0.5(\delta_1(p(1, h))\delta_d(y) + \delta_0(p(1, h))\delta_{1-d}(y)) \quad (4.177)$$

$$= \delta_{1-d}(y) \quad (4.178)$$

$$\neq \mathbb{P}_\alpha^{\mathbf{Y}_i|\mathbf{D}_i}(y|h, d) \quad i \neq 1 \quad (4.179)$$

Thus  $\mathbb{P}_C^{\mathbf{Y}|\mathbf{D}}$  is not causally contractible by Theorem 4.2.7.

However, given any finite permutation  $\rho : \mathbb{N} \rightarrow \mathbb{N}$

$$\mathbb{P}_\alpha^{\mathbf{Y}|\mathbf{I}}(y|q) = \sum_{h \in \overline{H}} \sum_{d \in \{0,1\}^{\mathbb{N}}} \prod_{i \in \mathbb{N}} \mathbb{P}_C^{\mathbf{Y}_i|\mathbf{I}_i\mathbf{D}_i\mathbf{H}}(y_i|q_i, d_i, h) \mathbb{P}_\alpha^{\mathbf{D}_i|\mathbf{I}_i\mathbf{H}}(d_i|q_i, h) \mathbb{P}_C^{\mathbf{H}}(h) \quad (4.180)$$

$$= \delta_{1-\alpha}(y_1) \delta_{(1)_{i \in \mathbb{N}}}(y_{>1}) \quad (4.181)$$

$$= \mathbb{P}_\alpha^{\mathbf{Y}|\mathbf{I}}(y|\rho^{-1}(q)) \quad (4.182)$$

$$= \mathbb{F}_\rho \mathbb{P}_\alpha^{\mathbf{Y}|\mathbf{I}}(y|q) \quad (4.183)$$

#### Example: body mass index

Given a sequential just-do model  $(\mathbb{P}_C, (\mathbf{B}, \mathbf{I}), \mathbf{Y})$  with  $\mathbf{B} := (\mathbf{B}_i)_{i \in M}$  representing body mass index of individual  $\mathbf{I}_i$  and  $\mathbf{Y} := (\mathbf{Y}_i)_{i \in M}$  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  $\mathbf{B}_i$  in the same way. Thus  $\mathbf{Y} \mathbf{I} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{C}|\mathbf{B}$  might generally be rejected, and so there may be no uniform conditional  $\mathbb{P}_C^{\mathbf{Y}|\mathbf{I}\mathbf{B}}$ . In this case,  $\mathbb{P}_C^{\mathbf{Y}|\mathbf{I}\mathbf{B}}$  cannot be causally contractible because it doesn't exist.

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

$$\mathbf{Y} \mathbf{B} \mathbf{I} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{C}|\mathbf{D} \quad (4.184)$$

$$\mathbf{Y} \mathbf{B} \mathbf{I} \perp\!\!\!\perp_{\mathbb{P}_C}^e \mathbf{D}|\mathbf{C} \quad (4.185)$$

and  $\mathbb{P}_C^{\mathbf{Y}\mathbf{B}|\mathbf{I}\mathbf{D}}$  causally contractible. Then by Theorem 4.3.4  $\mathbb{P}_C^{\mathbf{Y}|\mathbf{B}\mathbf{D}}$  is also causally contractible. In general, there may be some  $U \subset H$  such that for any  $h \in U$

$$\mathbb{P}_C^{\mathbf{Y}_i|\mathbf{B}_i\mathbf{D}_i\mathbf{H}}(y|b, d, h) = \mathbb{P}_C^{\mathbf{Y}_i|\mathbf{B}_i\mathbf{H}}(y|b, h) \quad (4.186)$$

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

Defining conditioning

So it may be possible to derive the fact that there is a repeatable response conditional  $\mathbb{P}_{C,H \in U}^{Y_i | H B_i}$  if  $H \in U$  is implied by available data, even if it is not assumed outright.

#### 4.4 Response conditionals with data-dependent actions

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 4.2.15 is generalised to sequential see-do models with the use of *probability combs*. Combs are a generalisation of conditional probabilities which can be used to capture a certain notion of data-dependent choices.

Because combs are a bit more complicated to work with than conditional probabilities, most of the previous results have not yet been generalised to the data-dependent case, if such a generalisation is possible.

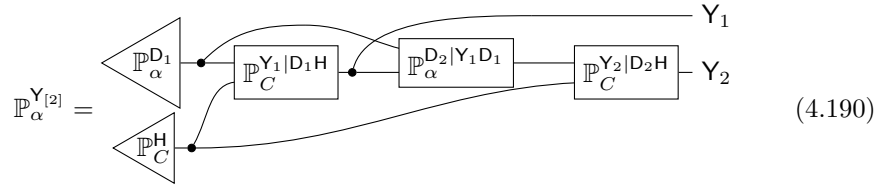
To begin with an example, consider a probability set  $(\mathbb{P}_C, D, Y)$  with  $D := (D_i)_{i \in \mathbb{N}}$  and  $Y := (Y_i)_{i \in \mathbb{N}}$  as usual, and take a subsequence  $(D_i, Y_i)_{i \in [2]}$  of length 2. Suppose  $\mathbb{P}_C$  features repeatable response conditionals in the sense that the following holds

$$Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{<i}, D_{<i}, C) | H D_i \quad \forall i \in \mathbb{N} \quad (4.187)$$

$$\wedge H \perp\!\!\!\perp_{\mathbb{P}_C}^e D C \quad (4.188)$$

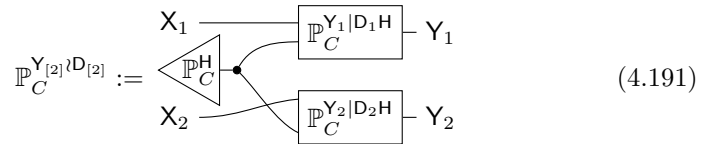
$$\wedge \mathbb{P}_C^{Y_i | H D_i} = \mathbb{P}_C^{Y_0 | H D_0} \quad \forall i \in \mathbb{N} \quad (4.189)$$

Then, for arbitrary  $\alpha \in C$



note that  $D_2$  depends on  $Y_1$  and  $D_1$ . Instead of multiplying by a distribution over  $(D_1, D_2)$ ,  $\mathbb{P}_\alpha^{D_2 | Y_1 D_1}$  has been “inserted” between the response conditionals  $\mathbb{P}_C^{Y_1 | D_1 H}$  and  $\mathbb{P}_C^{Y_2 | D_2 H}$ . A comb is a Markov kernel that yields a probability distribution when another Markov kernel of appropriate type is inserted in this manner.

Given  $\mathbb{P}_C^{Y_1 | D_1 H}$  and  $\mathbb{P}_C^{Y_2 | D_2 H}$ , define the comb



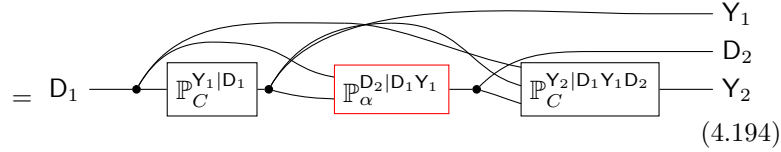
then  $\mathbb{P}_C^{Y_{[2]} \wr D_{[2]}}$  is causally contractible.  $\mathbb{P}_C^{Y_{[2]} \wr D_{[2]}}$  is *not* a uniform conditional probability; in general

$$\mathbb{P}_\alpha^{D_1 D_2} \mathbb{P}_C^{Y_{[2]} \wr D_{[2]}} \neq \mathbb{P}_\alpha^{Y_1 Y_2} \quad (4.192)$$

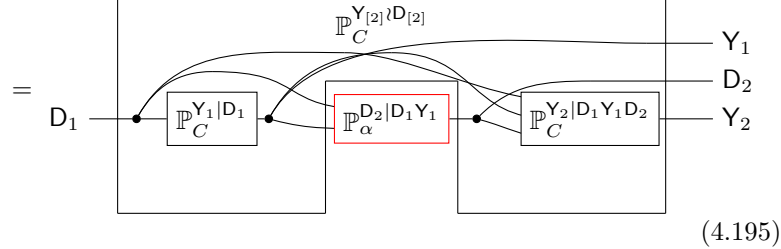
#### 4.4.1 Combs

Where uniform conditional probabilities map probability distributions to probability distributions via the semidirect product, 2-combs map conditional probabilities to conditional probabilities via an “insert” operation. More generally, higher order combs map lower order combs to lower order combs (where conditional probabilities are thought of as 1-combs and probability distributions as 0-combs). Graphically, the insert operation can be represented in with the following diagrams

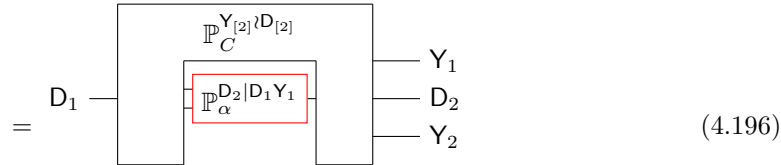
$$\mathbb{P}_\alpha^{Y_1 D_2 Y_2 | D_1} = \text{insert}(\mathbb{P}_\alpha^{D_2 | D_1 Y_1}, \mathbb{P}_C^{Y_{[2]} \wr D_{[2]}}) \quad (4.193)$$



(4.194)



(4.195)



(4.196)

While Equation 4.194 is a well-formed string diagram in the category of Markov kernels, Equation 4.196 is not. In the case that all the underlying sets are discrete, Equation 4.196 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.

**Definition 4.4.1** (Uniform  $n$ -Comb). Given a probability set  $\mathbb{P}_C$  with variables  $Y_i : \Omega \rightarrow Y$ ,  $D_i : \Omega \rightarrow D$  for  $i \in [n]$  and uniform conditional probabilities  $\{\mathbb{P}_C^{Y_i | D_{[i]} Y_{[i-1]}} | i \in [n]\}$ , the uniform  $n$ -comb  $\mathbb{P}_C^{Y_{[n]} \wr D_{[n]}} : D^n \rightarrow Y^n$  is the Markov

**Definition 4.4.5** (Sequential see-do model). A *sequential see-do model* is a triple  $(\mathbb{P}_C, D, Y)$  where  $\mathbb{P}_C$  is a probability set on  $(\Omega, \mathcal{F})$ ,  $D$  is a sequence of “inputs”  $D := (D_i)_{i \in \mathbb{N}}$  and  $Y$  is a corresponding sequence of “outputs”  $Y = (Y_i)_{i \in \mathbb{N}}$  where  $D_i : \Omega \rightarrow D$  and  $Y_i : \Omega \rightarrow Y$  and  $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^c C | (D_{[i]}, Y_{<i})$ .



**Theorem 4.4.6.** *Given a sequential see-do model  $(\mathbb{P}'_C, D', Y')$  on  $(\Omega, \mathcal{F})$ , then  $\mathbb{P}_C^{Y'D'}$  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  $H : \Omega \times H \rightarrow H$  such that  $Y_i \perp\!\!\!\perp_{\mathbb{P}'_C}^e (Y_{<i}, X_{<i}, C) | (X_i, H)$  and  $\mathbb{P}_C^{Y_i | X_i H} = \mathbb{P}_C^{Y_j | X_j H}$  for all  $i, j \in \mathbb{N}$  and  $H \perp\!\!\!\perp_{\mathbb{P}_C} (X, C)$ .*

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

$$\mathbb{P}_C^{Y_i | HD_i} = \mathbb{L} \quad (4.202)$$

and  $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{<i}, D_{<i}) | (D_i, H)$ . Thus

$$\mathbb{P}_C^{Y_i | HD_i Y_{<i} D_{<i}} = \mathbb{L} \otimes \text{erase}_{Y^{i-1} \times D^{i-1}} \quad (4.203)$$

and so

$$\mathbb{P}_C^{Y_i D} = \begin{array}{c} \triangleleft \mathbb{P}_C^H \\ \bullet \\ \boxed{\begin{array}{c} \text{---} \mathbb{L} \text{---} Y_i \\ \text{---} D_i \text{---} \\ i \in \mathbb{N} \end{array}} \end{array} \quad (4.204)$$

and so by Theorem 4.2.11,  $\mathbb{P}_C^{Y_i D}$  is causally contractible.

Only if: First, define the extension  $\mathbb{P}_C$ . From Theorem 4.2.11 and causal contractibility of  $\mathbb{P}_C^{Y'D'}$  there is some  $H, \mu \in \Delta(H)$  and  $\mathbb{L} : H \times D \rightarrow Y$  such that

$$\mathbb{P}_C^{Y'D'} = \begin{array}{c} \triangleleft \mu \\ \bullet \\ \boxed{\begin{array}{c} \text{---} \mathbb{L} \text{---} Y_i \\ \text{---} D_i \text{---} \\ i \in \mathbb{N} \end{array}} \end{array} \quad (4.205)$$

thus, by the definition of the comb insert operation

$$\mathbb{P}_\alpha^{D'_{[n]} Y'_{[n]}} = \mathbb{P}_\alpha^{D_1} \odot \text{insert}(\mathbb{P}_\alpha^{D'_{[2,n]} Y'_{[n-1]}}, \mathbb{P}_C^{Y'_{[n]} D'_{[n]}}) \quad (4.206)$$

Let

$$\mathbb{P}_C^{Y_i | HD_i} = \mathbb{L} \quad (4.207)$$

and let  $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{<i}, D_{<i}) | (D_i, H)$ , and for all  $\alpha$  set  $\mathbb{P}_\alpha^{W | DY} = \mathbb{P}_\alpha^{W' | D' Y'}$  for all  $W' : \Omega \rightarrow W$  and  $\mathbb{P}_\alpha^{D_i | Y_{<i} D_{<i}} = \mathbb{P}_\alpha^{D'_i | Y'_{<i} D'_{<i}}$ .

It remains to be shown that  $\mathbb{P}_\alpha^{DY} = \mathbb{P}_\alpha^{D'Y}$ .

By Equation 4.207 and  $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y_{<i}, D_{<i}) | (D_i, H)$ , it follows (for identical

reasons as Equation 4.204) that

$$\mathbb{P}_C^{Y_i D} = \begin{array}{c} \triangleleft \mathbb{P}_C^H \\ \text{---} \bullet \text{---} \boxed{\mathbb{L}} \text{---} Y_i \\ \text{---} D_i \text{---} \text{---} i \in \mathbb{N} \end{array} \quad (4.208)$$

$$= \begin{array}{c} \triangleleft \mu \\ \text{---} \bullet \text{---} \boxed{\mathbb{L}} \text{---} Y_i \\ \text{---} D_i \text{---} \text{---} i \in \mathbb{N} \end{array} \quad (4.209)$$

$$= \mathbb{P}_C^{Y'_i D'} \quad (4.210)$$

And so for all  $n \in \mathbb{N}$

$$\mathbb{P}_\alpha^{D_{[n]} Y_{[n]}} = \mathbb{P}_\alpha^{D_1} \odot \text{insert}(\mathbb{P}_\alpha^{D_{[2,n]} Y_{[n-1]}}, \mathbb{P}_C^{Y_{[n]} D_{[n]}}) \quad (4.211)$$

$$= \mathbb{P}_\alpha^{D_1} \odot \text{insert}(\mathbb{P}_\alpha^{D'_{[2,n]} Y'_{[n-1]}}, \mathbb{P}_C^{Y'_{[n]} D'_{[n]}}) \quad (4.212)$$

$$= \mathbb{P}_\alpha^{D'_{[n]} Y'_{[n]}} \quad (4.213)$$

□

In contrast to the data-independent case where causal contractibility of  $\mathbb{P}_C^{Y|X}$  implies the equivalence of all subsequence conditionals  $\mathbb{P}_C^{Y_A|X_A}$  for all equally sized  $A \subset \mathbb{N}$ , a causally contractible comb  $\mathbb{P}_C^{Y|D}$  does not generally imply that subsequence combs  $\mathbb{P}_C^{Y_A|D_A}$  and  $\mathbb{P}_C^{Y_B|D_B}$  are equivalent with  $|A| = |B|$ .

### 4.4.3 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  $\mathbb{P}_\alpha^{Y_{[2]}|D_{[2]}}$ . Then (assuming discrete sets)

$$\mathbb{P}_\alpha^{Y_{[2]}|D_{[2]}}(y_1, y_2 | d_1, d_2) = \mathbb{P}_\alpha^{Y_1|D_1}(y_1 | d_1) \mathbb{P}_\alpha^{Y_2|D_2}(y_2 | d_2) \quad (4.214)$$

$$= \frac{\mathbb{P}_\alpha^{Y_1|D_1}(y_1 | d_1) \mathbb{P}_\alpha^{D_2|Y_1 D_1}(d_2 | y_1, d_1) \mathbb{P}_\alpha^{Y_2|D_2}(y_2 | d_2)}{\mathbb{P}_\alpha^{D_2|Y_1 D_1}(d_2 | y_1, d_1)} \quad (4.215)$$

$$= \frac{\mathbb{P}_\alpha^{Y_{[2]}|D_2|D_1}(y_1, y_2, d_2 | d_1)}{\mathbb{P}_\alpha^{D_2|Y_1 D_1}(d_2 | y_1, d_1)} \quad (4.216)$$

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.

## 4.5 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  $\alpha$  (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 superscripter “ $o$ ” 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  $Y^o := (Y_i^o)_{i \in \mathbb{N}}$  is a sequence of observational variables,  $Y^v$  a sequence of interventional variables and  $Y^{o,v} := (Y_i^o, Y_i^v)_{i \in \mathbb{N}}$  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  $\mathbb{P}_C^{Y^{o,v} | X^{o,v} D^{o,v}}$  with  $D^v$  responsive to choices and  $X^v$  unresponsive (see Example 4.2.17). 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  $\mathbb{P}_C^{Y^{o,v} | U^{o,v} D^{o,v}}$ , where  $U^{o,v}$  is unobserved. That is, while  $U^{o,v}$  appears in the model, it is not associated with any measurement procedure. This model still asserts that  $(U_i^o, X_i^o, Y_i^o)$  triples are interchangeable with  $(U_i^v, X_i^v, Y_i^v)$  triples, but neither of these are measurement outcomes. On the other hand,  $(D_i^o, Y_i^o)$  pairs are not generally interchangeable with  $(D_i^v, Y_i^v)$ .

Consider models that satisfy causal contractibility with respect to  $\mathbb{P}_C^{Y^{o,v} | W^{o,v}}$ , where no comment is made about whether  $W^{o,v}$  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”<sup>1</sup>). More precisely, for any choice  $\alpha \in C$  and any consequence  $Y_i^v$ , there is a random subsequence  $Q$  of indices  $(1, 2, 3, \dots)$  such that the distribution  $\mathbb{P}_\alpha^{Y^{o,v}}$  is unchanged by permutations that only swap elements in the sequence  $(RVY_Q^o, Y_i^v)$ .

**Theorem 4.5.1.** *Given just-do model  $\mathbb{P}_C$  with  $\mathbb{P}_C^{Y^{o,v} | W^{o,v}}$  causally contractible,  $W$  finite and  $\mathbb{P}_C^{W^o | H}(w|h) > 0$  for all  $w, h$ , define  $q : W^\mathbb{N} \times W \rightarrow (\{*\} \cup \mathcal{P}(\mathbb{N}))$  by*

$$q : ((w_j^o)_\mathbb{N}, w_i^v) \mapsto \{j | w_j^o = w_i^v\} \quad (4.217)$$

<sup>1</sup>Ecclesiastes 1:9 reads “Everything that happens has happened before; nothing is new, nothing under the sun.” (noa, 1995)

and take  $Q := q \circ (W^o, W_i^v)$  for arbitrary  $i \in \mathbb{N}$ . For an index set  $U \in \mathbb{N}$  Take  $\text{swap}_U : Y^{\mathbb{N}} \times Y^{\mathbb{N}} \rightarrow Y^{\mathbb{N}} \times Y^{\mathbb{N}}$  to be an arbitrary finite swap that acts as the identity on all indices  $(j, x) \notin Q \times \{o\} \cup \{(i, v)\}$ . Then  $\mathbb{P}^{Y^o Y_i^v} \text{swap}_Q = \mathbb{P}^{Y^o Y_i^v}$ .

*Proof.* Note that for  $B_j \in \mathcal{W}$ , where  $\rho_q : \mathbb{N} \times \{i, v\} \rightarrow \mathbb{N} \times \{i, v\}$  is the permutation function associated with  $\text{swap}_q$

$$\mathbb{P}_\alpha^{W^o W_i^v} \text{swap}_Q \left( \bigotimes_{j \in \mathbb{N}} B_j \right) = \int_{W^{\mathbb{N}}} \int_{\mathcal{P}(\mathbb{N})} \prod_{k \notin q \times \{o\} \cup \{(i, v)\}} \delta_{w_k}(B_k) \prod_{l \in q \times \{o\} \cup \{(i, v)\}} \delta_{\rho_q(w_l)}(B_l) \mathbb{P}_\alpha^{Q|W^o W_i^v}(dq|w) \mathbb{P}_\alpha(dw) \quad (4.218)$$

$$= \int_{W^{\mathbb{N}}} \int_{\mathcal{P}(\mathbb{N})} \prod_{k \notin q \times \{o\} \cup \{(i, v)\}} \delta_{w_k}(B_k) \prod_{l \in q \times \{o\} \cup \{(i, v)\}} \delta_{w_l}(B_l) \mathbb{P}_\alpha^{Q|W^o W_i^v}(dq|w) \mathbb{P}_\alpha(dw) \quad (4.219)$$

$$= \mathbb{P}_\alpha^{W^o W_i^v} \quad (4.220)$$

where Eq. 4.219 follows from the fact that for every  $k, l \in q \times \{o\} \cup \{(i, v)\}$ ,  $w_k = w_l$ .

Thus for  $A \in \mathcal{Y}^{\mathbb{N}}$

$$\mathbb{P}_\alpha^{Y^o Y_i^v} \text{swap}_Q(A) = [\mathbb{P}_\alpha^{W^o W_i^v} \mathbb{P}_\alpha^{Y^o Y_i^v | Q W^o W_i^v} \text{swap}_Q](A) \quad (4.221)$$

$$= [\mathbb{P}_\alpha^{W^o W_i^v} \text{swap}_{Q^{-1}} \mathbb{P}_\alpha^{Y^o Y_i^v | W^o W_i^v} \text{swap}_Q](A) \quad (4.222)$$

$$= \mathbb{P}_\alpha^{Y^o Y_i^v} \quad (4.223)$$

Where Eq. 4.223 follows from causal contractibility of  $\mathbb{P}_\alpha^{Y^o Y_i^v | W^o W_i^v}$ .  $\square$

It also follows from Ecclesiastes' assumption and finite  $W$  that if some  $X_i^o$ ,  $Z_i^o$  are *deterministically* related given  $W$ , then  $\mathbb{P}_C^{Z_i^o | X_i^o}$  is causally contractible.

**Theorem 4.5.2.** *Given just-do model  $\mathbb{P}_C$  with  $\mathbb{P}_C^{X^{o,v} Z^{o,v} | W^{o,v}}$  causally contractible,  $W$  finite and  $\mathbb{P}_C^{W_i^o | H}(w|h) > 0$  for all  $w, h$ , if  $\mathbb{P}_C^{Z_0^o | X_0^o H}$  is deterministic then  $\mathbb{P}_C^{Z^{o,v} | X^{o,v}}$  is causally contractible.*

*Proof.* Because  $\mathbb{P}_\alpha^{W_0^o} \mathbb{P}_C^{Z_0^o | X_0^o W_0^o H}$  is deterministic, so is  $\mathbb{P}_C^{Z_0^o | X_0^o W_0^o H}$ .

Fix  $h \in H$ . Suppose there is some  $w, w' \in W$  such that

$$\mathbb{P}_C^{Z_0^o | X_0^o W_0^o H}(A|x, w, h) \neq \mathbb{P}_C^{Z_0^o | X_0^o W_0^o H}(A|x, w', h) \quad (4.224)$$

then, by determinism, we can assume without loss of generality

$$\mathbb{P}_C^{Z_0^o | X_0^o W_0^o H}(A|x, w, h) = 1 \quad (4.225)$$

$$\mathbb{P}_C^{Z_0^o | X_0^o W_0^o H}(A|x, w', h) = 0 \quad (4.226)$$

but  $W$  is finite and  $\mathbb{P}_C^{W_i^o | H}(w|h) > 0$  for all  $w$ , so there is some  $a > 0$  such that  $\mathbb{P}_C^{W_i^o | H}(w|h) \geq a$  for all  $w$ , and so

$$a \leq \sum_{w \in W} \mathbb{P}_C^{W_0^o | X_0^o H}(w|x, h) \mathbb{P}_C^{Z_0^o | X_0^o W_0^o H}(A|x, w, h) \leq 1 - a \quad (4.227)$$

contradicting determinism of  $\mathbb{P}_C^{Z_0^\circ | X_0^\circ H}$ .

Thus for all  $w, w'$

$$\mathbb{P}_C^{Z_0 | X_0 W_0 H}(A|x, w, h) = \mathbb{P}_C^{Z_0 | X_0 W_0 H}(A|x, w', h) \quad (4.228)$$

i.e.  $Z_0 \perp\!\!\!\perp_{\mathbb{P}_C}^e (W_0, C) | (X_0, H)$ . But then there is some  $\mathbb{P}_C^{Z_0 | X_0 H}$  such that

$$\mathbb{P}_C^{Z_0 | X_0 W_0 H} = \mathbb{P}_C^{Z_0 | X_0 H} \otimes \text{erase}_W \quad (4.229)$$

$$\implies \mathbb{P}_\alpha^{Z_i^v | X_i^v H} = \mathbb{P}_C^{Z_0 | X_0 H} \quad (4.230)$$

□

Theorem 4.5.2 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 occurs very rarely in observations. On the other hand, values of  $W$  rare in observations might, under some choices, become common.

**Example 4.5.3.** Say  $\mathbb{P}_C^{Z_i^\circ | X_i^\circ H}$  is *approximately deterministic* if  $\mathbb{P}_C^{Z_i^\circ | X_i^\circ H}(A|x, h) \in [0, \epsilon] \cup [1 - \epsilon, 1]$  for all  $A \in \mathcal{Z}$ ,  $x, h \in X \times H$ .

Take  $Z = X = W = H = \{0, 1\}$ . Set

$$\mathbb{P}_C^{Z_0 | X_0 W_0 H}(1|1, 1, 1) = 1 \quad (4.231)$$

$$\mathbb{P}_C^{Z_0 | X_0 W_0 H}(1|1, 0, 1) = 0 \quad (4.232)$$

and

$$\mathbb{P}_C^{W_0^\circ | H}(1|1) = 1 - \epsilon \quad (4.233)$$

then

$$\mathbb{P}_C^{Z_0 | X_0 H}(1|1, 1) = 1 - \epsilon \quad (4.234)$$

however, suppose there is some  $\alpha$  such that

$$\mathbb{P}_\alpha^{W_i^v | H}(1|1) = 0 \quad (4.235)$$

then

$$\mathbb{P}_\alpha^{Z_0 | X_0 H}(1|1, 1) = 0 \quad (4.236)$$

$$\neq \mathbb{P}_C^{Z_0 | X_0 H}(1|1, 1) \quad (4.237)$$



## Chapter 5

# See-do models, interventions and counterfactuals

### 5.1 How do see-do models relate to other approaches to causal inference?

- Review of approaches: CBN, CBN soft intervention, CBN fat-hand intervention, CBN noise intervention, SEM (Pearl/Heckman), PO unit model, PO population model, SWIG, Dawid decision theoretic model, Heckerman decision theoretic model, Rohde/Lattimore Bayesian model
- Focus on CBN, PO unit model, PO population model

### 5.2 Interpretations of the choice set

- Decisions or actions we could actually make - decision problem
- Idealised/hypothetical choices constrained by a set of causal relationships - interventions
- Suppositions - counterfactuals
- Further possibility - intervention  $\rightarrow$  decisions might be actuator randomisation

### 5.3 Causal Bayesian Networks as see-do models

- Definition of CBN, intervention set (recall: existence of disintegrations, decomposability)
- How interventions differ from decisions: no effect strength uncertainty, side effects, may be more interventions than what we actually know how to do

- Example: sets of CBNs and d-separation

## 5.4 Unit Potential Outcomes models

- Counterfactual random variables  $Y_x$  answer a question: "what would  $Y$  be supposing  $X$  was  $x$ ?"
- Proposed formalisation of suppositions: (....)
- Implies existence of counterfactual random variables
- Difference between suppositions and decisions: determinism, other conditions
- "3-player models": hypotheses, suppositions and interventions/decisions
- Error in key theorem of Rubin, Imbens (ignorability does not imply functional exchangeability)
- What can be represented by a 3 player model?
  - "1 of 2 counterfactuals": anything
  - "3 of 2 counterfactuals": very restrictive
  - "2 of 3 counterfactuals": Bell's theorem, counterfactual definiteness

This chapter is currently a disorganised cut and paste

The field of causal inference is additionally concerned with types of questions called "counterfactual" by Pearl. There is substantial theoretical interest in counterfactual questions, but counterfactual questions are much more rarely found in applications than interventional questions. Even though see-do models are motivated by the need to answer interventional questions, the theory developed here is surprisingly applicable to counterfactuals as well. In particular, the theory of see-do models offers explanations for three key features of counterfactual models:

- **Apparent absence of choices:** *Potential outcomes* models, which purportedly answer counterfactual questions, are standard statistical models *without choices* (Rubin, 2005)
- **Deterministic dependence on unobserved variables:** Counterfactual models involve *deterministic* dependence on unobserved variables (Pearl, 2009; Rubin, 2005; Richardson and Robins, 2013)
- **Residual dependence on observations:** Counterfactual questions depend on the given data *even if the joint distribution of this data is known*. For example, Pearl (2009) introduces a particular method for conditioning a known joint distribution on observations that he calls *abduction*



Potential outcomes models lack a notion of “choices” because there is a generic method to “add choices” to a potential outcomes model, which is implicitly used whenever potential outcomes models are used. Furthermore, we show that a see-do model induces a potential outcome model if and only if it is a model of *parallel choices*, and in this case the observed consequences depend deterministically on the unobserved potential outcomes in precisely the manner as given in Rubin (2005). Parallel choices can be roughly understood as models of sequences of experiments where an action can be chosen for each experiment, and with the special properties that repeating the same action deterministically yields the same consequence, and the consequences of a sequence of actions doesn’t depend on the order in which the actions are taken. That is, we show that the fundamental property of any “counterfactual” model is *deterministic reproducibility* and *action exchangeability*, and while these models may admit a “counterfactual” interpretation, they are fundamentally just a special class of see-do models.

But the proof is still in my notebook

Interestingly, it seems to be possible to construct a see-do model where the “hypothesis” is a quantum state, and quantum mechanics + locality seems to rule out parallel choices in such models in a manner similar to Bell’s theorem. “Seems to” because I haven’t actually proven any of these things.

The residual dependence on observations exhibited by counterfactual questions is a generic property of see-do models, and it is a particular property of *decision problems* are notable in that it is often

Where to discuss the connections to statistical decision theory?

See-do models are closely related to *statistical decision theory* introduced by Wald (1950) and elaborated by Savage (1954) after Wald’s death. See-do models equipped with a *utility function* induce a slightly generalised form of statistical decision problems, and the complete class theorem is applicable to these models.

A stylistic difference between see-do models and most other causal models is that see-do models explicitly represent both the observation model and the consequence model and their coupling, making them “two picture” causal models. Causal Bayesian Networks and Single World Intervention Graphs (Richardson and Robins, 2013) use “one picture” to represent the observation model and the consequence model. However, both of these approaches employ “graph mutilation”, so one picture on the page actually corresponds to many pictures when combined with the mutilation rules. For more on how these different types of models relate, see Section ?? . Lattimore and Rohde (2019)’s Bayesian causal inference employs two-picture causal models, as do “twin networks” (Pearl, 2009).

Sometimes we are interested in modelling situations where we can also make some choices that also affect the eventual consequences. For example, I might hypothesise  $H_1$ : the switch on the wall controls my light,  $H_2$ : the switch on the wall does not control my light. Then, given  $H_1$  I can choose to toggle the switch, and I will see my light turn on, or I can choose not to toggle the switch and I will

not see my light turn on. Given  $H_2$ , neither choice will result in a light turned on. Choices are clearly different to hypotheses: the choice I make depends on what I want to happen, while whether or not a hypothesis is true has no regard for my ambitions.

A “statistical model with choices” is simply a map  $\mathbb{T} : D \times H \rightarrow \Delta(\mathcal{E})$  for some set of choices  $D$ , hypotheses  $H$  and outcome space  $(E, \mathcal{E})$ . We can also distinguish two types of outcomes: *observations* which are given prior to a choice being made and *consequences* which happen after a choice is made. Observations cannot be affected by the choices made, while consequences are not subject to this restriction. That is, observations are what we might *see* before making a choice, which depends on the hypothesis alone, and if we are lucky we may be able to invert this dependence to learn something about the hypothesis from observations. On the other hand, the consequences of what we *do* depends jointly on the hypothesis and the choice we make and we judge which choices are more desirable on the basis of which consequences we expect them to produce.

What we are studying is a family of models that generalises of statistical models to include hypotheses, choices, observations and consequences. These models are referred to as *see-do models*. Hypotheses, observations, consequences and choices are not individually new ideas. *Statistical decision problems* (Wald, 1950; ?) extend statistical models with decisions and *losses*. Like consequences, losses depend on which choices are made. However, unlike consequences, losses must be ordered and reflect the preferences of a decision maker. *Influence diagrams* are directed graphs created to represent decision problems that feature “choice nodes”, “chance nodes” and “utility nodes”. An influence diagram may be associated with a particular probability distribution Nilsson and Lauritzen (2013) or with a set of probability distributions Dawid (2002).

See-do models have deep roots in decision theory. Decision theory asks, out of a set of available acts, which ones ought to be chosen. See-do models answer an intermediate question: out of a set of available acts, what are the consequences of each? This question is described by Pearl (2009) as an “interventional” question.

See-do models depend crucially on a set of choices  $D$ . While these models can obviously answer questions like “what is likely to happen if I choose  $d \in D$ ?”, this construction appears to rule out “causal” questions like “Does rain cause wet roads?”. We define a restricted idea of causation called *D-causation*. Roughly, if the roads get wet when it rains regardless of my choice of  $d \in D$ , then rain “*D*-causes” wet roads. *D*-causation is closely related to the idea *limited invariance* put forward by Heckerman and Shachter (1995).

### 5.4.1 D-causation

The choice set  $D$  is a primitive element of a see-do model. However, while we claim that see-do models are the basic objects studied in causal inference, so far we have no notion of “causation”. What we call *D-causation* is one such notion. It is called *D*-causation because it is a notion of causation that depends on the set of choices available. A similar idea, called *limited unresponsiveness*, is discussed extensively in the decision theoretic account of causation found in Heckerman and Shachter

(1995). The main difference is that see-do maps are fundamentally stochastic while Heckerman and Shachter work with “states” (approximately hypotheses in our terminology) that map decisions deterministically to consequences. In addition, while we define  $D$ -causation relative to a see-do map  $\mathbb{T}$ , Heckerman and Shachter define limited unresponsiveness with respect to *sets* of states.

Section ?? explores the difficulty of defining “objective causation” without reference to a set of choices.  $D$  need not be interpreted as the set of choices available to an agent, but however we want to interpret it, all existing examples of causal models seem to require this set.

See Section ?? for the definition of random variables in Kernel spaces.

One way to motivate the notion of  $D$ -causation is to observe that for many decision problems, I may wish to include a very large set of choices  $D$ . Suppose I aim to have my light switched on, and there is a switch that controls the light. Often, the relevant choices for such a problem would appear to be  $D_0 = \{\text{flip the switch, don't flip the switch}\}$ . However, this doesn't come close to exhausting the set of things I might choose to do, and I might wish to consider a larger set of possibilities. For simplicity's sake, suppose I have instead the following set of options:

$$D_1 := \{ \text{“walk to the switch and press it with my thumb”}, \\ \text{“trip over the lego on the floor, hop to the light switch and stab my finger at it”}, \\ \text{“stay in bed”} \}$$

If having the light turned on is all that matters, I could consider any acts in  $D_1$  to be equivalent if, in the end, the light switch ends up in the same position. In this case, I could say that the light switch position  $D_1$ -causes the state of the light. Subject to the assumption that the light switch position  $D_1$ -causes the state of the light, I can reduce my problem to one of choosing from  $D_0$  (noting that some choices correspond to mixtures of elements of  $D_0$ ).

If I consider an even larger set of possible acts  $D_2$ , I might not accept that the switch position  $D_2$ -causes the state of the light. Let  $D_2$  be the following acts:

$$D_2 := \{ \text{“walk to the switch and press it with my thumb”}, \\ \text{“trip over the lego on the floor, hop to the light switch and stab my finger at it”}, \\ \text{“stay in bed”}, \\ \text{“toggle the mains power, then flip the light switch”} \}$$

In this case, it would be unreasonable to suppose that all acts that left the light switch in the “on” position would also result in the light being “on”. Thus the switch does not  $D_2$ -cause the light to be on.

Formally,  $D$ -causation is defined in terms of conditional independence. Given a see-do model  $\mathbb{T} : H \times D \rightarrow \Delta(\mathcal{X} \otimes \mathcal{Y})$ , define the *consequence model*  $\mathbb{C} : H \times D \rightarrow \Delta(\mathcal{Y})$  as  $\mathbb{C} := \mathbb{T}^{\mathcal{Y}|\mathcal{H}D}$ .

**Definition 5.4.1** (*D-causation*). Given a hypothesis  $h \in H$  and a consequence model  $\mathbb{C} : H \times D \rightarrow \Delta(\mathcal{Y})$ , random variables  $Y_1 : Y \times D \rightarrow Y_1$ ,  $Y_2 : Y \times D \rightarrow Y_2$  and  $D : Y \times D \rightarrow D$  (defined the usual way),  $Y_1$  *D-causes*  $Y_2$  iff  $Y_2 \perp\!\!\!\perp_{\mathbb{C}} D | Y_1 H$ .

### 5.4.2 D-causation vs Limited Unresponsiveness

Heckerman and Shachter study deterministic “consequence models”. Furthermore, what we call hypotheses  $h \in H$ , Heckerman and Schachter call states  $s \in S$ . Heckerman and Shachter’s notion of causation is defined by *limited unresponsiveness* rather than *conditional independence*, which depends on a partition of states rather than a particular hypothesis.

**Definition 5.4.2** (Limited unresponsiveness). Given states  $S$ , deterministic consequence models  $\mathbb{C}_s : D \rightarrow \Delta(F)$  for each  $s \in A$  and a random variables  $Y_1 : F \rightarrow Y_1$ ,  $Y_2 : F \rightarrow Y_2$ ,  $Y_1$  is unresponsive to  $D$  in states limited by  $Y_2$  if  $\mathbb{C}_{(s,d)}^{Y_2|SD} = \mathbb{C}_{(s,d')}^{Y_2|SD} \implies \mathbb{C}_{(s,d)}^{Y_1|SD} = \mathbb{C}_{(s,d')}^{Y_1|SD}$  for all  $d, d' \in D$ ,  $s \in S$ . Write  $Y_1 \not\prec_{Y_2} D$

**Lemma 5.4.3** (Limited unresponsiveness implies *D-causation*). *For deterministic consequence models,  $Y_1 \not\prec_{Y_2} D$  implies  $Y_2$  D-causes  $Y_1$ .*

*Proof.* By the assumption of determinism, for each  $s \in S$  and  $d \in D$  there exists  $y_1(s, d)$  and  $y_2(s, d)$  such that  $\mathbb{C}_{s,d}^{Y_1 Y_2 | SD} = \delta_{y_1(s,d)} \otimes \delta_{y_2(s,d)}$ .

By the assumption of limited unresponsiveness, for all  $d, d'$  such that  $y_2(s, d) = y_2(s, d')$ ,  $y_1(s, d) = y_1(s, d')$  also. Define  $f : Y_2 \times S \rightarrow Y_1$  by  $(s, y_1) \mapsto y(s, [y_1(s, \cdot)]^{-1}(y_1(s, d)))$  where  $[y_1(s, \cdot)]^{-1}(a)$  is an arbitrary element of  $\{d | y_1(s, d) = a\}$ . For all  $s, d$ ,  $f(y_1(s, d), s) = y_2(s, d)$ . Define  $\mathbb{M} : Y_2 \times S \times D \rightarrow \Delta(\mathcal{Y}_1)$  by  $(y_2, s, d) \mapsto \delta_{f(y_2, s)}$ .  $\mathbb{M}$  is a version of  $\mathbb{C}^{Y_1 | Y_2, S, D}$  because, for all  $A \in \mathcal{Y}_2$ ,  $B \in \mathcal{Y}_1$ ,  $s \in S$ ,  $d \in D$ :

$$\mathbb{C}_{(s,d)}^{Y_2|SD} \Upsilon (\mathbb{M} \otimes \text{Id}) = \int_A \mathbb{M}(y'_2, d, s; B) d\delta_{y_2(s,d)}(y'_2) \quad (5.1)$$

$$= \int_A \delta_{f(y'_2, s)}(B) d\delta_{y_2(s,d)}(y'_2) \quad (5.2)$$

$$= \delta_{f(y_2(s,d), s)}(B) \delta_{y_2(s,d)}(A) \quad (5.3)$$

$$= \delta_{y_1(s,d)}(B) \delta_{y_2(s,d)}(A) \quad (5.4)$$

$$= \delta_{y_2(s,d)} \otimes \delta_{y_1(s,d)}(A \times B) \quad (5.5)$$

$\mathbb{M}$  is clearly constant in  $D$ . Therefore  $Y_1 \perp\!\!\!\perp_{\mathbb{C}} D | Y_2 S$ .  $\square$

define this

However, despite limited unresponsiveness implying *D-causation*, it does not imply *D-causation* in mixtures of states. Suppose  $D = \{0, 1\}$  where 1 stands for “toggle light switch” and 0 stands for “do nothing”. Suppose  $S = \{[0, 0], [0, 1], [1, 0], [1, 1]\}$  where  $[0, 0]$  represents “switch initially off, mains off” the other states generalise this in the obvious way. Finally,  $F \in \{0, 1\}$  is the final position of the switch and  $L \in \{0, 1\}$  is the final state of the light. We have

$$\mathbb{C}_{d,[i,m]}^{\text{LF|DS}} = \delta_{(d \text{ XOR } i) \text{ AND } m} \otimes \delta_{(d \text{ XOR } i) \text{ AND } m} \quad (5.6)$$

Within states  $[0, 0]$  and  $[1, 0]$ , the light is always off, so  $F = a \implies L = 0$  for any  $a$ . In states  $[0, 1]$  and  $[1, 1]$ ,  $F = 1 \implies L = 1$  and  $F = 0 \implies L = 0$ . Thus  $L \not\prec_F D$ . However, suppose we take a mixture of consequence models:

$$\mathbb{C}_\gamma = \frac{1}{4}\mathbb{C}_{\cdot,[0,0]} + \frac{1}{4}\mathbb{C}_{\cdot,[0,1]} + \frac{1}{2}\mathbb{C}_{\cdot,[1,1]} \quad (5.7)$$

$$\mathbb{C}_\gamma^{\text{FL|D}} = \frac{1}{4} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} + \frac{1}{4} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (5.8)$$

Then

$$[1, 0]\mathbb{C}_\gamma^{\text{FL|D}} = \frac{1}{4}[0, 1] \otimes [1, 0] + \frac{1}{4}[0, 1] \otimes [0, 1] + \frac{1}{2}[1, 0] \otimes [1, 0] \quad (5.9)$$

$$[1, 0]\vee(\mathbb{C}_\gamma^{\text{F|D}} \otimes \mathbb{C}_\gamma^{\text{L|D}}) = (\frac{1}{2}[0, 1] + \frac{1}{2}[1, 0]) \otimes (\frac{1}{4}[0, 1] + \frac{3}{4}[1, 0]) \quad (5.10)$$

$$\implies [1, 0]\mathbb{C}_\gamma^{\text{FL|D}} \neq [1, 0]\vee(\mathbb{C}_\gamma^{\text{F|D}} \otimes \mathbb{C}_\gamma^{\text{L|D}}) \quad (5.11)$$

Thus under the prior  $\gamma$ ,  $F$  does not  $D$ -cause  $L$  even though  $F$   $D$ -causes  $L$  in all states  $S$ . The definition of  $D$ -causation was motivated by the idea that we could reduce a difficult decision problem with a large set  $D$  to a simpler problem with a smaller “effective” set of decisions by exploiting conditional independence. Even if  $X$   $D$ -causes  $Y$  in every  $H \in S$ ,  $X$  does not necessarily  $D$ -cause  $Y$  in mixtures of states in  $S$ . For this reason, we do not say that  $X$   $D$ -causes  $Y$  in  $S$  if  $X$   $D$ -causes  $Y$  in every  $H \in S$ , and in this way we differ substantially from Heckerman and Shachter (1995).

define this

Instead, we simply extend the definition of  $D$ -causation to mixtures of hypotheses: if  $\gamma \in \Delta(H)$  is a mixture of hypotheses, define  $\mathbb{C}_\gamma := (\gamma \otimes \text{Id})\mathbb{C}$ . Then  $X$   $D$ -causes  $Y$  relative to  $\gamma$  iff  $Y \perp\!\!\!\perp_{\mathbb{C}_\gamma} D|X$ .

Theorem 5.4.4 shows that under some conditions,  $D$ -causation can hold for arbitrary mixtures over subsets of the hypothesis class  $H$ .

**Theorem 5.4.4** (Universal  $D$ -causation). *If  $X \perp\!\!\!\perp H|D$  for all  $H, H' \in S \subset H$  and  $X$   $D$ -causes  $Y$  in all  $H \in S$ , then  $X$   $D$ -causes  $Y$  with respect to all mixed consequence models  $\mathbb{C}_\gamma$  for all  $\gamma \in \Delta(H)$  with  $\gamma(S) = 1$ .*

*Proof.* For  $\gamma \in \Delta(H)$ , define the mixture

$$\mathbb{C}_\gamma := \begin{array}{c} \triangleleft \gamma \\ \text{D} \text{---} \boxed{\mathbb{C}} \text{---} F \end{array} \quad (5.12)$$

Because  $\mathbb{C}_H^{\text{X|D}} = \mathbb{C}_{H'}^{\text{X|D}}$  for all  $H, H' \in H$ , we have

$$(5.13)$$

Also

$$(5.14)$$

$$(5.15)$$

$$(5.16)$$

$$(5.17)$$

$$(5.18)$$

$$(5.19)$$

Equation 5.19 establishes that  $(\gamma \otimes \mathbf{Id}_X \otimes \dagger_D)C^{Y|XH}$  is a version of  $C_\gamma^{Y|XD}$ , and thus  $Y \perp\!\!\!\perp_{C_\gamma} D | X$ .

This can also be derived from the semi-graphoid rules:

$$H \perp\!\!\!\perp D \wedge H \perp\!\!\!\perp X | D \implies H \perp\!\!\!\perp XD \quad (5.20)$$

$$\implies H \perp\!\!\!\perp D | X \quad (5.21)$$

$$D \perp\!\!\!\perp H | X \wedge D \perp\!\!\!\perp Y | XH \implies D \perp\!\!\!\perp Y | X \quad (5.22)$$

$$\implies Y \perp\!\!\!\perp D | X \quad (5.23)$$

□

### 5.4.3 Properties of D-causation

If  $X$  D-causes  $Y$  relative to  $\mathbb{C}_H$ , then the following holds:

$$\mathbb{C}_H^{X|D} = D - \boxed{\mathbb{C}^{X|D}} - \boxed{\mathbb{C}^{Y|X}} - Y \quad (5.24)$$

This follows from version (2) of Definition ??:

$$\mathbb{C}_H^{X|D} = D - \boxed{\mathbb{C}^{X|D}} - \boxed{\mathbb{C}^{Y|XD}} - Y \quad (5.25)$$

$$= D - \boxed{\mathbb{C}^{X|D}} - \boxed{\mathbb{C}^{Y|X}} - Y \quad (5.26)$$

$$= D - \boxed{\mathbb{C}^{X|D}} - \boxed{\mathbb{C}^{Y|X}} - Y \quad (5.27)$$

D-causation is not transitive: if  $X$  D-causes  $Y$  and  $Y$  D-causes  $Z$  then  $X$  doesn't necessarily D-cause  $Z$ .

Pearl's “front door adjustment” and general identification results make use of composing “sub-consequence-kernels” like this. Show, if possible, that Pearl's “sub-consequence-kernels” obey  $D$ -causation like relations

Does this “weak  $D$ -causation” respect mixing under the same conditions as regular  $D$ -causation?

### 5.4.4 Decision sequences and parallel decisions

Just as observations  $X$  can be a sequence of random variables  $X_1, X_2, \dots$ ,  $D$  can be a sequence of “sub-choices”  $D_1, D_2, \dots$ . Note that by positing such a sequence there is no requirement that  $D_1$  comes “before”  $D_2$  in any particular sense.

## 5.5 Existence of counterfactuals

I'm struggling with how to explain this well.

“Counterfactual” or “potential outcomes” models in the causal inference literature are consequence models where choices can be considered in *parallel*.

Before defining parallel choices, we will consider a “counterfactual model” without parallel choices. Consider the following definitions, first from Pearl (2009) pg. 203-204. I have preserved his notation, including not using any special fonts for things called “variables” because this term is used interchangeably with “sets of variables” and using special fonts for variables might give the impression that these should be treated as different things while using special fonts for sets of variables is inconsistent with my usual notation.

The real solution here is that Pearl’s “variable sets” are actually “coupled variables”, see Definition ??, but I’d rather not change his definitions if I can avoid it

put the following inside a quote environment somehow, the regular quote environment fails due to too much markup

““

**Definition 7.1.1 (Causal Model)** A causal model is a triple  $M = \langle U, V, F \rangle$ , where:

- (i)  $U$  is a set of *background* variables, (also called *exogenous*), that are determined by factors outside the model;
- (ii)  $V$  is a set  $\{V_1, V_2, \dots, V_n\}$  of variables, called *endogenous*, that are determined by variables in the model – that is, variables in  $U \cup V$ ;
- (iii)  $F$  is a set of functions  $\{f_1, f_2, \dots, f_n\}$  such that each  $f_i$  is a mapping from (the respective domains of)  $U_i \cup PA_i$  to  $V_i$ , where  $U_i \subseteq U$  and  $PA_i \subseteq V \setminus V_i$  and the entire set  $F$  forms a mapping from  $U$  to  $V$ . In other words, each  $f_i$  in

$$v_i = f_i(pa_i, u_i), \quad i \in 1, \dots, n,$$

assigns a value to  $V_i$  that depends on (the values of) a select set of variables in  $V \cup U$ , and the entire set  $F$  has a unique solution  $V(u)$ .

**Definition 7.1.2 (Submodel)** Let  $M$  be a causal model,  $X$  a set of variables in  $V$ , and  $x$  a particular realization of  $X$ . A submodel  $M_x$  of  $M$  is the causal model

$$M_x = \{U, V, F_x\},$$

where

$$F_x = \{f_i : V_i \notin X\} \cup \{X = x\}.$$

**Definition 7.1.3 (Effect of Action)** Let  $M$  be a causal model,  $X$  a set of variables in  $V$ , and  $x$  a particular realization of  $X$ . The effect of action  $do(X = x)$  on  $M$  is given by the submodel  $M_x$

**Definition 7.1.4 (Potential Response)** Let  $X$  and  $Y$  be two subsets of variables in  $V$ . The potential response of  $Y$  to action  $do(X = x)$ , denoted  $Y_x(u)$ , is the solution for  $Y$  of the set of equations  $F_x$ , that is,  $Y_x(u) = Y_{M_x}(u)$ .



**Definition 7.1.6 (Probabilistic Causal Model)** A probabilistic causal model is a pair  $\langle M, P(u) \rangle$ , where  $M$  is a causal model and  $P(u)$  is a probability function defined over the domain of  $U$ . ”

Implicitly, Definition 7.1.3 proposes a set of “actions” that have “effects” given by  $M_x$ . It’s not entirely clear what this set of actions should be – the definition seems to suggest that there is an action for each “realization” of each variable in  $V$ , which would imply that the set of actions corresponds to the range of  $V$ . For the following discussion, we will call the set of actions  $D$ , whatever it actually contains (we have deliberately chosen to use the same letter as we use to represent choices or actions in see-do models).

Given  $D$ , Definition 7.1.3 appears to define a function  $h : \mathcal{M} \times D \rightarrow \mathcal{M}$ , where  $\mathcal{M}$  is the space of causal models with background variables  $U$  and endogenous variables  $V$ , such that for  $M \in \mathcal{M}$ ,  $do(X = x) \in D$ ,  $h(M, do(X = x)) = M_x$ .

Definition 7.1.4 then appears to define a function  $Y(\cdot) : D \times U \rightarrow Y$  (distinct from  $Y$ , which appears to be a function  $U \rightarrow \text{something}$ ) and calls  $Y(\cdot)$  the “potential response”. We could always consider the variable  $V := \bigotimes_{i \in [n]} V_i$  and define the “total potential response”  $\mathbf{g} := V(\cdot)$ , which captures the potential responses of any subset of variables in  $V$ .

From this, we might surmise that in the Pearlean view, it is necessary that a “counterfactual” or “potential response” model has a probability measure  $P$  on background variables  $U$ , a set of actions  $D$  and a *deterministic* potential response function  $\mathbf{g} : D \times U \rightarrow V$ .

Pearl’s model also features a second deterministic function  $\mathbf{f} : U \rightarrow Y$ , and  $G$  is derived from  $F$  via the equation modifications permitted by  $D$ . It is straightforward to show that an arbitrary function  $\mathbf{f} : U \rightarrow Y$  can be constructed from Pearl’s set of functions  $f_i$ , and if  $D$  may modify the set  $F$  arbitrarily, then it appears that  $\mathbf{g}$  can in principle be an arbitrary function  $D \times U \rightarrow Y$  (though many possible choices would be quite unusual).

Pearl’s counterfactual model seems to essentially be a deterministic map  $\mathbf{g} : D \times U \rightarrow V$  along with a probability measure  $P$  on  $U$ . Putting these together and marginalising over  $U$  (as we might expect we want to do with “background variables”) simply yields a consequence map  $D \rightarrow \Delta(\mathcal{V})$ , which doesn’t seem to have any special counterfactual properties.

In order to pose counterfactual questions, Pearl introduces the idea of holding  $U$  fixed:

““

**Definition 7.1.5 (Counterfactual)** Let  $X$  and  $Y$  be two subsets of variables in  $V$ . The counterfactual sentence “ $Y$  would be  $y$  (in situation  $u$ ), had  $X$  been  $x$ ” is interpreted as the equality  $Y_x(u) = y$ , with  $Y_x(u)$  being the potential response of  $Y$  to  $X = x$ . ”

Holding  $U$  fixed allows SCM counterfactual models to answer questions

about what would have happened if we had taken different actions given the same background context. For example, we can compare  $Y_x(u)$  with  $Y_{x'}(u)$  and interpret the comparison as telling us what would have happened in the same situation  $u$  if we did  $x$  and, at the same time, what would happen if we did  $x'$ . It is the ability to consider different actions “in exactly the same situation” that makes these models “counterfactual”.

One obvious question is: does  $\mathbf{g}$  have to be deterministic? While SCMs are defined in terms of deterministic functions with noise arguments, it's not clear that this is a necessary feature of counterfactual models. If  $\mathbf{g}$  were properly stochastic, what is the problem with considering  $\mathbf{g}(x, u)$  and  $\mathbf{g}(x', u)$  to represent what would happen in a fixed situation  $u$  if I did  $x$  and if I did  $x'$  respectively? In fact, a nondeterministic  $\mathbf{g}$  arguably fails to capture a key intuition of taking actions “in exactly the same situation”. If I want to know the result of doing action  $x$  and, in exactly the same situation, the result of doing action  $x$ , then one might intuitively think that the result should always be *deterministically the same*. This property, which we call *deterministic reproducibility*, does not hold if we consider a nondeterministic potential response map  $\mathbf{g}$ .

This idea of doing  $x$  and, in the same situation, doing  $x$  doesn't render very well in English. Furthermore, even though deterministic reproducibility seems to be an important property of counterfactual SCMs, they don't help very much to elucidate the idea. “If I take action  $x$  in situation  $U$  I get  $V_x(u)$  and if I take action  $x$  in situation  $U$  I get  $V_x(u)$ ” is just a redundant repetition. It seems that we want some way to express the idea of having two copies of  $V_x(u)$  or, more generally, having multiple copies of a potential response function in such a way that we can make comparisons between their results.

The idea that we need *can* be clearly expressed with a see-do model.

## Chapter 6

# Other causal modelling frameworks

### References

- Holy Bible : Contemporary English Version*. New York : American Bible Society, [1995] 1995, 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 Validity. We thank Esther Duflo for her leadership on the handbook and for extensive comments on earlier drafts. Chassang and Snowberg gratefully acknowledge the support of NSF grant SES-1156154. In Abhijit Vinayak Banerjee and Esther Duflo, editors, *Handbook of Economic Field Experiments*, volume 1 of *Handbook of Field Experiments*, pages 141–174. North-Holland, January 2017. doi: 10.1016/bs.hefe.2016.08.005. URL <https://www.sciencedirect.com/science/article/pii/S2214658X16300071>.
- 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>.
- 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. P. Dawid. Influence Diagrams for Causal Modelling and Inference. *International Statistical Review*, 70(2):161–189, 2002. ISSN 1751-5823. doi: 10.1111/j.1751-5823.2002.tb00354.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2002.tb00354.x>.
- 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](https://doi.org/10.1007/978-1-4612-0919-5_10).
- 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. Interditions, 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>.
- 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>.
- D. Heckerman and R. Shachter. Decision-Theoretic Foundations for Causal Reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, December 1995. ISSN 1076-9757. doi: 10.1613/jair.202. URL <https://www.jair.org/index.php/jair/article/view/10151>.
- M. A. Hernán and S. L. Taubman. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity*, 32(S3):S8–S14, August 2008. ISSN 1476-5497. doi: 10.1038/ijo.2008.82. URL <https://www.nature.com/articles/ijo200882>.
- Miguel A. Hernán 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=rJNeXOgdbR>.
- 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 Bojaczek 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](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>.
- 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.
- Finnian Lattimore and David Rohde. Replacing the do-calculus with Bayes rule. *arXiv:1906.07125 [cs, stat]*, December 2019. URL <http://arxiv.org/abs/1906.07125>. arXiv: 1906.07125.
- 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>.
- 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](https://doi.org/10.1007/978-3-7091-6045-9_31).
- Dennis Nilsson and Steffen L. Lauritzen. Evaluating Influence Diagrams using LIMIDs. *arXiv:1301.3881 [cs]*, January 2013. URL <http://arxiv.org/abs/1301.3881>. arXiv: 1301.3881.
- 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 cigarette-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>.
- 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](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 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](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=tv&paperid=4710&option\\_lang=eng](http://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=tv&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](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:**