Causal Statistical Decision Theory|What are interventions?

David Johnston

April 22, 2022

Contents

Chapter 1

Introduction

1.1 Theories of causal inference

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

The question of a causal link between smoking and cancer is a very important one to many different people. Individuals who enjoy smoking (or think they might) may wish to avoid smoking if cigarettes pose a severe health risk, so they are interested in knowing whether or not it is so. Additionally, some may desire reassurance that their habit is not too risky, whether or not this is true. Potential and actual investors in cigarette manufacturers may see health concerns as a barrier to adoption, and also may personally want to avoid supporting products that harm many people. Like smokers, such people might have some interest in knowing the truth of this question, and a separate interest in hearing that cigarettes are not too risky, whether or not this is true. Governments and organisations with a responsibility for public health may see

themselves as having responsibility to discourage smoking as much as possible if smoking is severely detrimental to health. The costs and benefits of poor decisions about smoking are large: 8 million annual deaths are attributed to cigarette-caused cancer and vascular disease in 2018(World Health Organisation, 2018) while global cigarette sales were estimated at US$711 billion in 2020 (Statista, 2020) (a figure which might be substantially larger if cigarettes were not widely believed to be harmful).

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

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

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

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

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

Theories of causation help us to answer causal questions, which means that before we have any theory, we already have causal questions we want to answer. If potential outcomes notation and causal graphical models had never been invented there would still be just as many people who want to the answer to questions something like "does smoking causes cancer?", even if on-one could

say what exactly they meant by "causes" and even if many people actually want answers to slightly different questions. Theories exist to serve our need for transparent answers to causal questions.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Chapter 2

Technical Prerequisites

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

Less likely to be familiar to readers is the string diagram notation we use to represent probabilistic functions. This is a notation created for reasoning about abstract Markov categories, and is somewhat different to existing graphical languages. The main difference is that in our notation wires represent variables and boxes (which are like nodes in directed acyclic graphs) represent probabilistic functions. Standard directed acyclic graphs annotate nodes with variable names and represent probabilistic functions implicitly. The advantage of explicitly representing probabilistic functions is that we can write equations involving graphics. It is introduced in Section 2.3.

We also extend the theory of probability to a theory of probability sets, which we introduce in Section 3.1.2. 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 3.1.2 for relevant generalisations to probability sets. Two key ideas introduced here are *uniform conditional probability*, similar but not identical to conditional probability, and *extended conditional independence* as introduced by Constantinou and Dawid (2017), similar but not identical to regular conditional independence.

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

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

2.1 Conventions

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

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

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

2.2 Probability Theory

2.2.1 Standard Probability Theory

σ-algebras

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

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

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

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

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

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

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

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

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

Probability measures and Markov kernels

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

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

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

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

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

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

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

Common probability measures and Markov kernels

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

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

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

Proof.

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

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

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

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

□

Variables, conditionals and marginals

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

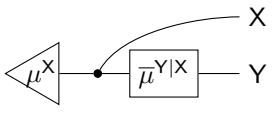
Definition 2.2.14 (Sequence of variables)**.** Given a measurable space $(\Omega, \mathcal{F})$ and two variables $\mathsf{X} : (\Omega, \mathcal{F}) \to (X, \mathcal{X})$, $\mathsf{Y} : (\Omega, \mathcal{F}) \to (Y, \mathcal{Y})$, $(\mathsf{X}, \mathsf{Y}) : \Omega \to X \times Y$ is the variable $\omega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

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

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

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

$$\iff \tag{2.6}$$

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

(2.7)

Markov kernel product notation

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

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

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

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

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

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

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

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

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

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

for all $x \in X$.

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

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

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

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

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

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

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

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

$$\begin{aligned} \mu\mathbb{F}_{\mathsf{X}}(A) &= \int_\Omega \delta_{\mathsf{X}(\omega)}(A)\mathrm{d}\mu(\omega) && (2.13)\\ &= \int_{\mathsf{X}^{-1}(\omega)} \mathrm{d}\mu(\omega) && (2.14)\\ &= \mu^{\mathsf{X}}(A) && (2.15) \end{aligned}$$

□

Semidirect product

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

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

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

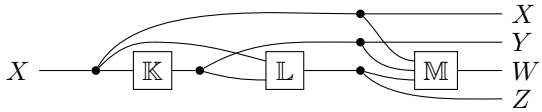
Lemma 2.2.23 (Semidirect product is associative)**.** *Given* $\mathbb{K} : X \to Y$, $\mathbb{L} : Y \times X \to Z$ *and* $\mathbb{M} : Z \times Y \times X \to W$

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

$$\tag{2.18}$$

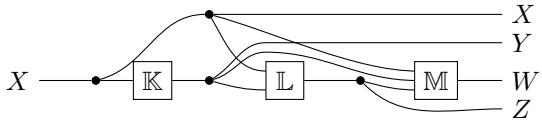
Proof.

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



(2.19)

$=$



(2.20)

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

□

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

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

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

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

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

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

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

2.3 String Diagrams

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

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

2.3.1 Elements of string diagrams

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

$\mathbb{K} :=$ (2.23)

$\mu :=$ (2.24)

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

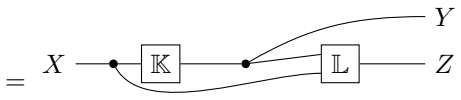
$$\mathbb{LM} := \; X \;\boxed{\mathbb{K}}\!\!-\!\!\boxed{\mathbb{M}}\; Z \tag{2.25}$$

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

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

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

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

$$=$$ 

(2.28)

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

$$\mathrm{Id}^X := \; X \longrightarrow X \tag{2.29}$$

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

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

$$= \; X \times Y \longrightarrow X \times Y \tag{2.31}$$

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

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

$$\equiv \tag{2.33}$$

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

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

2.3.2 Special maps

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

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

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

$$\mathrm{del}_X := \quad —* X \tag{2.36}$$

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

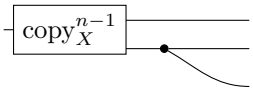
$$\mathrm{swap}_{X,Y} := \quad \text{  } \tag{2.37}$$

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

$$\mathrm{copy}_X := \quad X \text{ } \begin{matrix} X \\ X \end{matrix} \tag{2.38}$$

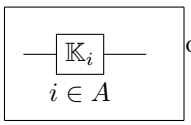
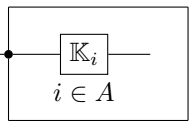
Definition 2.3.5 (n-fold copy map)**.** The n-fold copy map $\mathrm{copy}_X^n : X \to\!\!\!\!\bullet\, X^n$ is given by the recursive definition

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

$$\mathrm{copy}_X^n = \quad \text{  } \qquad n > 1 \tag{2.40}$$

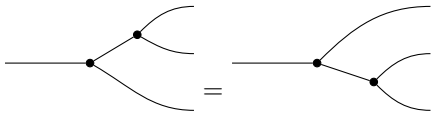
Plates In a string diagram, a plate that is annotated $i \in A$ means the tensor product of the $|A|$ elements that appear inside the plate. A wire crossing from outside a plate boundary to the inside of a plate indicates an $|A|$-fold copy map, which we indicate by placing a dot on the plate boundary.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

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

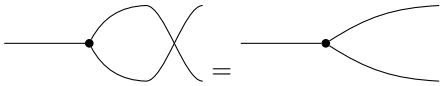
2.3.3 Commutative comonoid axioms

Diagrams in Markov categories satisfy the commutative comonoid axioms.



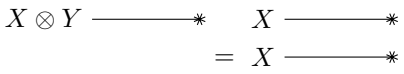
(2.42)

(2.43)

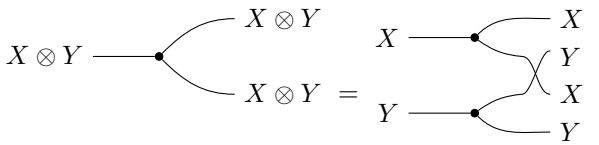


(2.44)

as well as compatibility with the monoidal structure

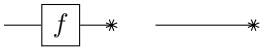


(2.45)



(2.46)

and the naturality of *del*, which means that

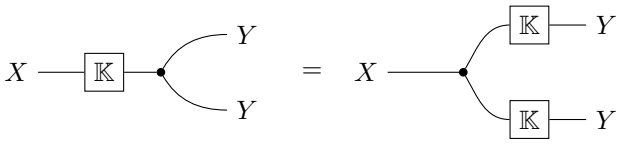


(2.47)

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* $\mathbb{K} : X \nrightarrow Y$



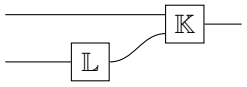
(2.48)

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



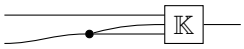
$$= (\mathrm{id}_X \otimes \mathbb{L})\mathbb{K} \tag{2.49}$$

$$[(\mathrm{id}_X \otimes \mathbb{L})\mathbb{K}](A|x,w) = \int_Y \int_X \mathbb{K}(A|x',y')\mathbb{L}(\mathrm{d}y'|w)\delta_x(\mathrm{d}x') \tag{2.50}$$

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

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

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



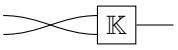
$$= (\mathrm{id}_X \otimes \mathrm{copy}_Y)\mathbb{K} \tag{2.52}$$

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

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

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

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



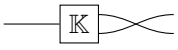
$$= \mathrm{swap}_{YX}\mathbb{K} \tag{2.55}$$

$$(\mathrm{swap}_{YX}\mathbb{K})(A|y,x) = \int_{X\times Y} \mathbb{K}(A|x',y')\delta_y(\mathrm{d}y')\delta_x(\mathrm{d}x') \tag{2.56}$$

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

The swap map before a kernel switches the input arguments.

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



$$= \mathbb{K}\mathrm{swap}_{YZ} \tag{2.58}$$

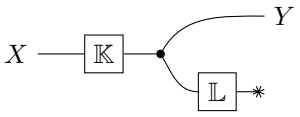
$$(\mathbb{K}\mathrm{swap}_{YZ})(A \times B|x) = \int_{Y\times Z} \delta_y(B)\delta_z(A)\mathbb{K}(\mathrm{d}y \times \mathrm{d}z|x) \tag{2.59}$$

$$= \int_{B\times A} \mathbb{K}(\mathrm{d}y \times \mathrm{d}z|x) \tag{2.60}$$

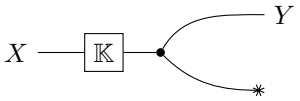
$$= \mathbb{K}(B \times A|x) \tag{2.61}$$

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

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



$$= \tag{2.63}$$



by Eq. 2.47

$$= \tag{2.64}$$

by Eq. 2.43

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

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

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

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

2.4 Probability Sets

A probability set is a set of probability measures. This section establishes a number of useful properties of conditional probability with respect to probability sets. Unlike conditional probability with respect to a probability space, conditional probabilities don't always exist for probability sets. Where they do, however, they are almost surely unique and we can marginalise and disintegrate them to obtain other conditional probabilities with respect to the same probability set.

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

Given a probability set $\mathbb{P}_C$, we define marginal and conditional probabilities as probability measures and Markov kernels that satisfy Definitions 2.2.15 and 2.2.16 respectively for *all* base measures in $\mathbb{P}_C$. There are generally multiple Markov kernels that satisfy the properties of a conditional probability with respect to a probability set, and this definition ensures that marginal and conditional probabilities are "almost surely" unique (Definition 2.4.7) with respect to probability sets.

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

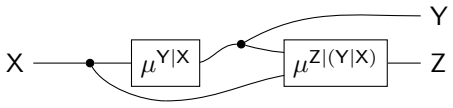
Definition 2.4.3 (Uniform conditional distribution)**.** Given a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$ and a probability set $\mathbb{P}_C$, a uniform conditional distribution $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ is any Markov kernel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

Given a conditional distribution $\mu^{\mathsf{ZY}|\mathsf{X}}$ we can define a higher order conditional $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$, which is a version of $\mu^{\mathsf{Z}|\mathsf{XY}}$. This is useful because uniform conditionals don't always exist, but we can use higher order conditionals to show that if a probability set $\mathbb{P}_C$ has a uniform conditional $\mathbb{P}_C^{\mathsf{ZY}|\mathsf{X}}$ then it also has a uniform conditional $\mathbb{P}_C^{\mathsf{Z}|\mathsf{XY}}$ (Theorems 2.4.26 and 2.4.27). Given $\mu^{\mathsf{XY}|\mathsf{Z}}$ and $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ standard measurable, it has recently been proven that a higher order conditional $\mu^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ exists Bogachev and Malofeev (2020), Theorem 3.5.

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

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

$$\iff \tag{2.69}$$



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

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

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

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

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

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

2.4.1 Almost sure equality

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

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

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

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

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

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

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

□

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

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

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

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

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

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

□

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

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

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

$$\mathbb{P}_\alpha^{\mathsf{XYZ}} = \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_\alpha^{\mathsf{YZ}|\mathsf{X}} \tag{2.80}$$

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

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

□

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

$$\mathbb{P}^{\mathsf{Y}|\mathsf{WX}} \overset{\mathbb{P}}{\cong} \quad \mathsf{W} \text{ —— } \boxed{\mathbb{K}} \text{ —— } \mathsf{Y} \quad ; \quad \mathsf{X} \text{ ——} * \tag{2.83}$$

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

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

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

$$\mathbb{P}^{\mathsf{Y}|\mathsf{WX}} \overset{\mathbb{P}}{\cong} \quad \mathsf{W} \text{ —— } \boxed{\mathbb{K}} \text{ —— } \mathsf{Y} \quad ; \quad \mathsf{X} \text{ ——} * \tag{2.85}$$

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

Proof. See Cho and Jacobs (2019). □

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

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

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

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

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

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

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

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

$$\iff \tag{2.88}$$

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

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

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

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

$$\iff \tag{2.91}$$

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

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

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

The following forms of these properties are often used here:

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

2.4.3 Examples

Example 2.4.18 (Choice variable)**.** Suppose we have a decision procedure $\mathcal{S}_C := \{\mathcal{S}_\alpha | \alpha \in C\}$ that consists of a measurement procedure for each element of a denumerable set of choices C. Each measurement procedure $\mathcal{S}_\alpha$ is modeled by a probability distribution $\mathbb{P}_\alpha$ on a shared sample space $(\Omega, \mathcal{F})$ such that we have an observable "choice" variable $(\mathsf{D}, \mathsf{D} \circ \mathcal{S}_\alpha)$ where $\mathsf{D} \circ \mathcal{S}_\alpha$ always yields α.

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

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

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

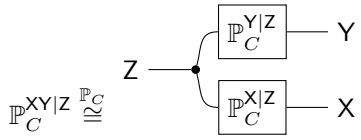
$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{D}} = \mathbb{P}_\alpha^{\mathsf{Y}} \tag{2.94}$$

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

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

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

Theorem 2.4.19 (Uniform conditional independence representation)**.** *Given a probability set $\mathbb{P}_C$ with a uniform conditional probability $\mathbb{P}_C^{\mathsf{XY}|\mathsf{Z}}$,*



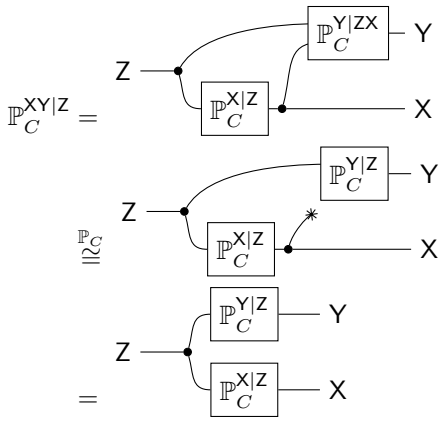
(2.96)

$$\iff \tag{2.97}$$

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

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

Proof. If: By Theorem 2.4.27

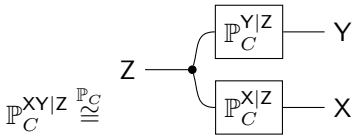


(2.99)

(2.100)

(2.101)

Only if: Suppose



(2.102)

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

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

then

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

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

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

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

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

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 "$\mathrm{do}(\mathsf{X} = x)$" specify a probability set by a collection of uniform conditional probabilities on variables other than X, along with marginal probabilities of X. Specifically:

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

The qualification "Y is a causal variable" is usually not an explicit condition for causal Bayesian networks, but it is an important one. For example, $2\mathsf{X}$ is not equal to X, but we cannot define a causal Bayesian network where both X and $2\mathsf{X}$ are causal variables, see Example 2.4.23.

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

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

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

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

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

We introduce the notion of *valid distributions* and *valid conditionals*. The key result here is: probability sets defined by collections of recursive valid conditionals and distributions are nonempty. While we suspect this condition is often satisfied by causal models in practice, we offer one example in the literature where it apparently is not. The problem of whether a probability set is valid is analogous to the problem of whether a probability distribution satisfying a collection of constraints exists discussed in Vorobev (1962). As that work shows, there are many questions of this nature that can be asked and that are not addressed by the criterion of validity.

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

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

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

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

Definition 2.4.22 (Maximal probability set)**.** Given $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$, $\mathsf{Y} : \Omega \to Y$ and a $\mathsf{Y}|\mathsf{X}$-valid conditional probability $\mathbb{L} : X \twoheadrightarrow Y$ the maximal probability set $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}[M]}$ associated with $\mathbb{L}$ is the probability set such that for all $\mathbb{P}_\alpha \in \mathbb{P}_C$, $\mathbb{L}$ is a version of $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}$.

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

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

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

Note that validity is not a necessary condition for a conditional to define a non-empty probability set. The intuition for this is: if we have some $\mathbb{K} : X \twoheadrightarrow Y$,

$\mathbb{K}$ might be an invalid $\mathsf{Y}|\mathsf{X}$ conditional on all of X, but might be valid on some subset of X, and so we might have some probability model $\mathbb{P}$ that assigns measure 0 to the bad parts of X such that $\mathbb{K}$ is a version of $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$. On the other hand, if we want to take the product of $\mathbb{K}$ with arbitrary valid X probabilities, then the validity of $\mathbb{K}$ is necessary (Theorem **??**).

Example 2.4.23. Body mass index is defined as a person's weight divided by the square of their height. Suppose we have a measurement process $\mathcal{S} = (\mathcal{W}, \mathcal{H})$ and $\mathcal{B} = \frac{\mathcal{W}}{\mathcal{H}^2}$ - i.e. we figure out someone's body mass index first by measuring both their height and weight, and then passing the result through a function that divides the second by the square of the first. Thus, given the random variables W, H modelling $\mathcal{W}, \mathcal{H}$, $\mathcal{B}$ is the function given by $\mathsf{B} = \frac{\mathsf{W}}{\mathsf{H}^2}$.

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

$$\mathbb{P}_\square^{\mathsf{B}|\mathsf{WHC}} = \begin{array}{ccc} \mathsf{H} & \longrightarrow\!\!* & \\ \mathsf{C} & \text{———————} & \mathsf{B} \\ \mathsf{W} & \longrightarrow\!\!* & \end{array} \tag{2.110}$$

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

$$(\mathsf{B}, \mathsf{W}, \mathsf{H}) \bowtie \{(x, w, h)\} = \{\omega | (\mathsf{W}, \mathsf{H})(\omega) = (w, h), \mathsf{B}(\omega) = \frac{w}{h^2}\} \tag{2.111}$$

$$= \emptyset \tag{2.112}$$

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

Validity rules out conditional probabilities like 2.110. 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 2.110 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 2.4.24 (Conditional pushforward)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_C$ with conditional $\mathbb{P}_C^{\mathsf{X}|\mathsf{Y}}$ such that $\mathsf{Z} = f \circ \mathsf{Y}$ for some $f : Y \to Z$. Then there exists a conditional probability $\mathbb{P}_C^{\mathsf{Z}|\mathsf{X}} = \mathbb{P}_C^{\mathsf{Y}|\mathsf{X}} \mathbb{F}_f$.*

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

$$\mathbb{P}_\alpha^{\mathsf{XZ}} = \mathbb{P}_\alpha^{\mathsf{XY}} \mathbb{F}_{\mathrm{id}_X \otimes f} \tag{2.113}$$

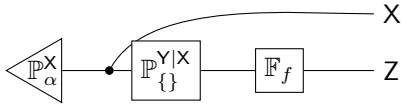
Note also that for all $A \in \mathcal{X}$, $B \in \mathcal{Z}$, $x \in X$, $y \in Y$:

$$\begin{aligned} \mathbb{F}_{\mathrm{id}_X \otimes f}(A \times B | x, y) &= \delta_x(A) \delta_{f(y)}(B) && (2.114)\\ &= \mathbb{F}_{\mathrm{id}_X}(A|x) \otimes \mathbb{F}_f(B|y) && (2.115)\\ \implies \mathbb{F}_{\mathrm{id}_X \otimes f} &= \mathbb{F}_{\mathrm{id}_X} \otimes \mathbb{F}_f && (2.116) \end{aligned}$$

Thus

$$\mathbb{P}_\alpha^{\mathsf{XZ}} = (\mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}) \mathbb{F}_{\mathrm{id}_X} \otimes \mathbb{F}_f \tag{2.117}$$

$$= \tag{2.118}$$



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

Theorem 2.4.25 (Existence of regular conditionals)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$ with Y standard measurable and a probability model $\mathbb{P}_\alpha$ on $(\Omega, \mathcal{F})$. Then there exists a conditional $\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}$.*

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

Theorem 2.4.26 (Existence of higher order valid conditionals with respect to probability sets)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_C$ with regular conditional $\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}$ and Y and Z standard measurable. Then there exists a regular $\mathbb{P}_C^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$.*

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

$$n(f^{-1}(y)|x, y) = 1 \tag{2.119}$$

$$m(\mathsf{Y}^{-1}(A) \cap B | x) = \int_A n(B|x, y) m\mathbb{F}_f(dy|x) \forall A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z} \tag{2.120}$$

In particular, $\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}$ is a Borel measurable map $X \twoheadrightarrow Y \times Z$. Thus equation 2.120 implies for all $A \in \mathcal{Y}, B \in \mathcal{Y} \times \mathcal{Z}$

$$\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}(\mathsf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x,y)\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}\mathbb{F}_f(dy|x) \tag{2.121}$$

$$= \int_A n(B|x,y)\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}(dy|x) \tag{2.122}$$

Where Equation 2.122 follows from Lemma 2.4.24.

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

$$\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}(\mathsf{Y}^{-1}(A) \cap B|x) = \int_A n(B|x,y)\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{X}}(dy|x) \tag{2.123}$$

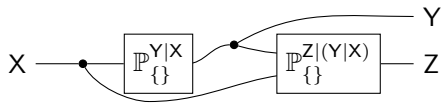
which implies n is a version of $\mathbb{P}_C^{\mathsf{YZ}|(\mathsf{Y}|\mathsf{X})}$. By Lemma 2.4.24, $n\mathbb{F}_f$ is a version of $\mathbb{P}_C^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$. □

We might be motivated to ask whether the higher order conditionals in Theorem 2.4.26 can be chosen to be valid. Despite Lemma **??** 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 2.119 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.

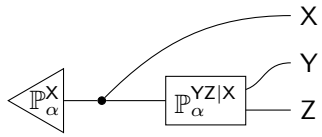
Theorem 2.4.27 (Higher order conditionals)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_C$ with conditional $\mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}$. Then $\mathbb{P}_C^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}$ is a version of $\mathbb{P}_C^{\mathsf{Z}|\mathsf{YX}}$*

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

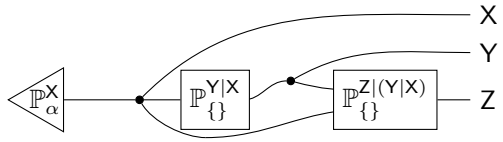


$$\mathbb{P}^{\mathsf{YZ}|\mathsf{X}}_\alpha = \tag{2.124}$$

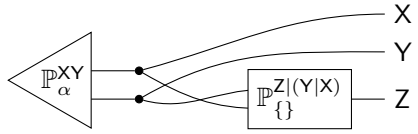
$$\implies \mathbb{P}^{\mathsf{XYZ}}_\alpha = \tag{2.125}$$



$$= \tag{2.126}$$



$$= \tag{2.127}$$

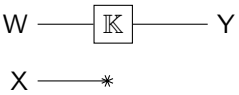


Thus $\mathbb{P}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}_C$ is a version of $\mathbb{P}^{\mathsf{Z}|\mathsf{YX}}_\alpha$ for all α and hence also a version of $\mathbb{P}^{\mathsf{Z}|\mathsf{YX}}_C$. □

Theorem 2.4.28. *Given probability gap model $\mathbb{P}_C$, X, Y, Z such that $\mathbb{P}^{\mathsf{Z}|\mathsf{YX}}_C$ exists, $\mathbb{P}^{\mathsf{Z}|\mathsf{Y}}_C$ exists iff $\mathsf{Z} \perp\!\!\!\perp_{\mathbb{P}_C} \mathsf{X}|\mathsf{Y}$.*

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

$$\mathbb{P}^{\mathsf{Y}|\mathsf{WX}}_\alpha = \tag{2.128}$$



□

Theorem 2.4.29 (Valid higher order conditionals)**.** *Suppose we have a sample space $(\Omega, \mathcal{F})$, variables $\mathsf{X} : \Omega \to X$ and $\mathsf{Y} : \Omega \to Y$, $\mathsf{Z} : \Omega \to Z$ and a probability set $\mathbb{P}_C$ with regular conditional $\mathbb{P}^{\mathsf{YZ}|\mathsf{X}}_C$, Y discrete and Z standard measurable. Then there exists a valid regular $\mathbb{P}^{\mathsf{Z}|\mathsf{XY}}_C$.*

Proof. By Theorem 2.4.26, we have a higher order conditional $\mathbb{P}^{\mathsf{Z}|(\mathsf{Y}|\mathsf{X})}_C$ which, by Theorem 2.4.27 is also a version of $\mathbb{P}^{\mathsf{Z}|\mathsf{XY}}_C$.

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

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

$$\forall A \in \mathcal{Z}, x \in X, y \in Y : \mathbb{P}_C^{\mathsf{YZ}|\mathsf{X}}(\{y\} \times A|x) = 0 \implies (\mathbb{Q}(A|x,y) = 0) \vee ((\mathsf{X},\mathsf{Y}) \bowtie \{(x,y)\} = \emptyset) \tag{2.129}$$

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

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

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, many theorists of causal inference signal an awareness that much of their analysis is ultimately motivated by problems involving a choice among options, even if such problems are not always directly addressed. Section 3.1 reviews briefly the role of decision problems according to theorists of causal inference and sets out the basic scheme by which probability sets are used to model problems of this type.

The reasons we provide for focusing on probability sets are not rigorous. The strongest motivation for this choice is *convention*: this chapter shows how many varieties of decision theory induce probability set models, and Chapter **??** shows how many causal inference frameworks also induce probability set models. Some decision theories examined in this chapter seek to justify their modelling choices by suggesting axioms for rational theories of decision under uncertainty, and deriving a particular type of model from these axioms. However, despite the various attempts at axiomatisation, the nature of such a theory is contested – there is no clear standard among the variety of theories surveyed here, nor among other proposed theories not considered in this work. This work is not

trying to resolve this dispute. However, modelling choices must still be made and it is still reasonable to ask what is being assumed by making these choices, so Section 3.2 provides an overview of several major decision theories along with their axiomatisations.

Section 3.2 describes in particular detail the connections between *statistical decision theory* (Wald, 1950) and probability set models of decision problems. There are two reasons for this: firstly, 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 **??** where we consider the kinds of probability set models induced by other causal reasoning frameworks.

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

3.1 What is the point of causal inference?

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

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

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

3. *Counterfactual*: informally, "questions concerning imagining alternate worlds";

formally defined as queries that can be answered by a structural causal model (SCM)

As with theories of rational decision making under uncertainty, adjudicating whether this is indeed a sound basis for causal inference is not the purpose of this thesis. The approach to causal inference presented here is motivated by the informal version of rung 2, rather than the informal version of rung 3 (the formal theory is founded on probability sets, not CBNs or SCMs). It is easy to doubt that rung 3 is the right informal starting point for a theory of causal inference: the argument that the ladder is hierarchical depends on the formal characterisation of rungs in terms of questions that can be posed *with particular kinds of models*. This gets priorities backwards; the point is to find models that are fit for answering important questions, not to find questions that can be asked of important models.

Furthermore, the informal questions that characterise rung 3 can be posed in language more reminiscent of rungs 2. Rung 2 is characterised, informally, by questions like "if I take aspirin, will my headache go away?" while rung 3 is characterised by questions like "if I had taken the aspirin, would I still have a headache?". However, the second question could be posed as "if I wave a magic wand and transport myself to a world where I have taken aspiring, would I still have a headache?". Such an intervention is fanciful, but it nevertheless sounds like an intervention. Thus, perhaps, one can characterise rung 3 by questions concerning interventions that may or may not be fanciful.

Rung 2 questions are far more prominent in the causal inference literature than rung 3. Rubin (2005) argues that causal inference often to 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. Problems that involve comparing different choices on the basis of their consequences are an important class of problems. They are worth better understanding, even if this work does not shed any light on counterfactual questions.

3.1.1 Modelling decision problems

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

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

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

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

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

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

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

3.1.2 Formal definitions

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

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

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

Definition 3.1.2 (Sharp model)**.** Given a decision problem (C, Ω, u), a *sharp model* is a function $C \twoheadrightarrow \Omega$.

Definition 3.1.3 (Vague model)**.** Given a decision problem (C, Ω, u), a *vague model* is a function $C \times H \twoheadrightarrow \Omega$ for some hypothesis set H.

Both sharp models and vague models have probability sets induced by the range of the model.

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

3.2 Representation theorems for decision problems

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

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

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

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

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

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

This definition is meant to correspond to the common sense idea of having preferences over some set of things, where $\succ$ can be read as "strictly better than", $\prec$ read as "strictly worse than" and $\sim$ read as "as good as". Given any

two things from the set, I can say which one I prefer, or if I prefer neither (and all of these are mutually exclusive). If I prefer a to a' then I think a' is worse than a. Furthermore, if I prefer a to a' and a' to a'' then I prefer a to a''.

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

3.2.1 von Neumann-Morgenstern utility

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

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

3.2.2 Savage decision theory

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

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

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

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

Proof. Savage (1954) □

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

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

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

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

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

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

Then

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

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

□

Savage axioms

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.2.3 Jeffrey's decision theory

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

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

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

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

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

Proof. Bolker (1966)) □

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

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

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

Where μ and des *are as in Theorem 3.2.4*

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

Bolker axioms

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

A1: $\preceq$ is a complete preference relation

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

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

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

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

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

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

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

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

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

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

A5 is an axiom of continuity.

3.2.4 Causal decision theory

Causal decision theory was developed after both Jeffrey's and Savage's theory. A number of authors Lewis (1981); Skyrms (1982) felt that Jeffrey's theory erred by treating the consequences of a choice as an "ordinary conditional probability". Lewis (1981) suggested that causal decision theory can be used to evaluate choices when we are given a set Ω of consequences over which preferences are known, a set C of choices and a set H of dependency hypotheses (the letters

have been changed to match usage in this work; in the original the consequences were called S, the choices A and the dependency hypotheses H). Choices are then evaluated according to the causal decision rule. We have taken the liberty to state Lewis' rule in the language of the present work.

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

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

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

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

Dependency hypotheses are quite an important idea in causal reasoning. Together Lewis' decision rule connect the theory of probability sets with *statistical decision theory*, as Section 3.2.5 will show. Chapter 4 goes into considerable detail concerning the question of when probability sets support certain types of dependency hypothesis. While they are typically not explicitly represented in common frameworks for causal inference, Chapter **??** 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 is a tuple $(X, H, D, l, \mathbb{P}_\cdot)$ where $(X, \mathcal{X})$ is a set of outcomes, $(H, \mathcal{H})$ is a set of hypotheses, $(D, \mathcal{D})$ is a set of decisions, $l : D \times H \to \mathbb{R}$ is a loss function and $\mathbb{P}_\cdot : H \to\!\!\!\!\triangleright\ X$ is a Markov kernel from hypotheses to to outcomes.

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

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

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 rule for assessing the desirability of a choice *relative to a particular hypothesis*, and considers a number of different rules for turning this "intermediate assessment" into a final assessment of the available choices.

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

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

SDT offers two decision rules that can be used to determine a best decision function when the hypothesis is unknown. The first is to choose the decision function minimising the *Bayes risk*, which requires a prior over hypotheses. The second decision rule is to choose the decision rule minimising the maximum risk, and this rule does not require a prior.

Definition 3.2.10 (Bayes risk)**.** Given a statistical decision problem $(X, H, D, l, \mathbb{P}_\cdot)$, decision functions C and a prior $\mu \in \Delta(\mathcal{H})$, the *Bayes risk* is

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

$$= \int_X \int_D l(d, h)\mathbb{D}_\alpha(\mathrm{d}d|f)\mathbb{P}_h(\mathrm{d}f)\mu(\mathrm{d}h) \tag{3.11}$$

Definition 3.2.11 (Minimax rule)**.** Given a statistical decision problem $(X, H, D, l, \mathbb{P}_\cdot)$ and decision functions C, a *minimax decision function* is any decision function $\mathbb{D}_\alpha$ satisfying

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

From consequences to statistical decision problems

To create a model *with consequences* that induces the SDT risk, we need to make a few assumptions. Recall that the basic model for a decision problem is either a Markov kernel $\mathbb{P}_\cdot : C \nrightarrow \Omega$ or $\mathbb{P}_{\cdot,\cdot} : C \times H \nrightarrow \Omega$ for some set of hypotheses H, and that a utility function u is also available. Here, Ω represents both observations (if any are taken) and consequences of decisions. Statistical decision problems correspond to a particular instance of this general scheme. In particular, statistical decision problems are associated with probability functions in which:

- The set of "choices" is identified with the set of decision rules
- There is some $\mathsf{X} : \Omega \to X$ representing the observations made before any decision rule is enacted
  - Given any hypothesis, observations are independent of the choice of decision rule
- There are consequences $\mathsf{Y} : \Omega \to Y$ and decisions $\mathsf{D} : \Omega \to D$
  - Given a hypothesis and a decision, the consequences Y are independent of both the choice of decision rule and the observations
  - Given a decision function, the decision D is independent of the hypothesis
- The utility $u : Y \to \mathbb{R}$ is a function of the consequences

We call models satisfying these requirements *see-do models*

Definition 3.2.12 (See-do model)**.** A probability set model of a statistical decision problem, or a *see-do model* for short, is a tuple $(\mathbb{P}_{C\times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D})$ where $\mathbb{P}_{C\times H}$ is a probability set indexed by elements of $C \times H$ on $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$ are the observations, $\mathsf{Y} : \Omega \to Y$ are the consequences and $\mathsf{D} : \Omega \to D$ are the decisions. $\mathbb{P}_{C\times H}$ must observe the following conditional independences:

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

$$\mathsf{Y} \perp\!\!\!\perp^{e}_{\mathbb{P}_{C\times H}} (\mathsf{X}, \mathsf{C})|(\mathsf{D}, \mathsf{H}) \tag{3.14}$$

$$\mathsf{D} \perp\!\!\!\perp^{e}_{\mathbb{P}_{C\times H}} \mathsf{H}|\mathsf{C} \tag{3.15}$$

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

In order to produce a statistical decision problem, we also need a loss. We identify the required loss with the negative expected utility, conditional on a particular decision.

Definition 3.2.13 (Induced loss)**.** Given a see-do model $(\mathbb{P}_{C\times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D})$ and a utility $u : Y \to \mathbb{R}$, the induced loss $l : D \times H \to \mathbb{R}$ is defined as

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

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

The probability set model induces a set of decision functions: for each $\alpha \in C$, there is a probability distribution $\mathbb{P}^{\mathsf{D}|\mathsf{X}}_{\alpha}$.

Theorem 3.2.14 (Induced risk)**.** *Given a see-do model $(\mathbb{P}_{C\times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D})$ along with a utility $u : Y \to \mathbb{R}$, the expected utility for each choice of decision rule $\alpha \in C$ and hypothesis $h \in H$ pair is equal to the negative risk of the decision rule and hypothesis pair.*

Proof. The expected utility given α and h is

$$\begin{aligned}
\int_Y u(y)\mathbb{P}^{\mathsf{Y}}_{\alpha,h}(\mathrm{d}y) &= \int_Y\int_D\int_X u(y)\mathbb{P}^{\mathsf{Y}|\mathsf{D}\mathsf{X}}_{\alpha,h}(\mathrm{d}y|d,x)\mathbb{P}^{\mathsf{D}|\mathsf{X}}_{\alpha,h}(\mathrm{d}d|x)\mathbb{P}^{\mathsf{X}}_{\alpha,h}(\mathrm{d}x) && (3.17)\\
&= \int_X\int_D\int_Y u(y)\mathbb{P}^{\mathsf{Y}|\mathsf{D}}_{\alpha,h}(\mathrm{d}y|d)\mathbb{P}^{\mathsf{D}|\mathsf{X}}_{\alpha,h}(\mathrm{d}d|x)\mathbb{P}^{\mathsf{X}}_{\alpha,h}(\mathrm{d}x) && (3.18)\\
&= \int_X\int_D\int_Y u(y)\mathbb{P}^{\mathsf{Y}|\mathsf{D}}_{C\times\{h\}}(\mathrm{d}y|d)\mathbb{P}^{\mathsf{D}|\mathsf{X}}_{\{\alpha\}\times H}(\mathrm{d}d|x)\mathbb{P}^{\mathsf{X}}_{C\times\{h\}}(\mathrm{d}x) && (3.19)\\
&= -\int_D\int_X l(d,h)\mathbb{P}^{\mathsf{D}|\mathsf{X}}_{\{\alpha\}\times H}(\mathrm{d}d|x)\mathbb{P}^{\mathsf{X}}_{C\times\{h\}}(\mathrm{d}x) && (3.20)\\
&= -R(\mathbb{P}^{\mathsf{D}|\mathsf{X}}_{\{\alpha\}\times H}, h) && (3.21)
\end{aligned}$$

where Equation 3.27 follows from $\mathsf{Y} \perp\!\!\!\perp^{e}_{\mathbb{P}_{C\times H}} (\mathsf{X},\mathsf{C})|(\mathsf{D},\mathsf{H})$, the uniform conditional $\mathbb{P}^{\mathsf{D}|\mathsf{X}}_{\{\alpha\}\times H}$ exists due to $\mathsf{D} \perp\!\!\!\perp^{e}_{\mathbb{P}_{C\times H}} \mathsf{H}|\mathsf{C}$ and the uniform conditional $\mathbb{P}^{\mathsf{X}}_{C\times\{h\}}$ exists due to $\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_{C\times H}} \mathsf{C}|\mathsf{H}$. □

We can also define a "Bayesian probability set model", which is a probability set model $\mathbb{P}_C$ along with a prior $\mathbb{P}^{\mathsf{H}}_C$ over hypotheses. In this case, the expected utility of a decision rule is equal to the Bayes risk.

Definition 3.2.15 (Bayesian see-do model)**.** A Bayesian see-do model is a tuple $(\mathbb{P}_C, \mathsf{X}, \mathsf{Y}, \mathsf{D}, \mathsf{H})$ where $\mathbb{P}_C$ is a probability set on $(\Omega, \mathcal{F})$, $\mathsf{X} : \Omega \to X$ are the observations, $\mathsf{Y} : \Omega \to Y$ are the consequences, $\mathsf{D} : \Omega \to D$ are the decisions and $\mathsf{H} : \Omega \to H$ is the hypothesis. $\mathbb{P}_C$ must observe the following conditional independences:

$$\mathsf{X} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{C}|\mathsf{H} \tag{3.22}$$
$$\mathsf{Y} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} (\mathsf{X},\mathsf{C})|(\mathsf{D},\mathsf{H}) \tag{3.23}$$
$$\mathsf{D} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{H}|\mathsf{C} \tag{3.24}$$
$$\mathsf{H} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{C} \tag{3.25}$$

Theorem 3.2.16 (Induced Bayes risk)**.** *Given a Bayesian see-do model* $(\mathbb{P}_C, \mathsf{X}, \mathsf{Y}, \mathsf{D}, \mathsf{H})$ *of a statistical decision problem along with a utility* $u : Y \to \mathbb{R}$*, the expected utility for each choice of decision rule* $\alpha \in C$ *is equal to the negative Bayes risk of that decision rule relative to the prior* $\mathbb{P}_C^{\mathsf{H}}$*.*

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

$$\int_Y u(y)\mathbb{P}_\alpha^{\mathsf{Y}}(\mathrm{d}y) = \int_Y \int_D \int_X \int_H u(y)\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{DXH}}(\mathrm{d}y|d,x,h)\mathbb{P}_\alpha^{\mathsf{D}|\mathsf{XH}}(\mathrm{d}d|x,h)\mathbb{P}_\alpha^{\mathsf{X}|\mathsf{H}}(\mathrm{d}x|h)\mathbb{P}_\alpha^{\mathsf{H}}(\mathrm{d}h) \tag{3.26}$$

$$= \int_X \int_D \int_Y \int_H u(y)\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{DH}}(\mathrm{d}y|d,h)\mathbb{P}_\alpha^{\mathsf{D}|\mathsf{X}}(\mathrm{d}d|x)\mathbb{P}_\alpha^{\mathsf{X}|\mathsf{H}}(\mathrm{d}x|h)\mathbb{P}_\alpha^{\mathsf{H}}(\mathrm{d}h) \tag{3.27}$$

$$= \int_X \int_D \int_Y \int_H u(y)\mathbb{P}_C^{\mathsf{Y}|\mathsf{DH}}(\mathrm{d}y|d,h)\mathbb{P}_\alpha^{\mathsf{D}|\mathsf{X}}(\mathrm{d}d|x)\mathbb{P}_C^{\mathsf{X}|\mathsf{H}}(\mathrm{d}x|h)\mathbb{P}_C^{\mathsf{H}}(\mathrm{d}h) \tag{3.28}$$

$$= -\int_D \int_X \int_H l(d,h)\mathbb{P}_\alpha^{\mathsf{D}|\mathsf{X}}(\mathrm{d}d|x)\mathbb{P}_C^{\mathsf{X}|\mathsf{H}}(\mathrm{d}x|h)\mathbb{P}_C^{\mathsf{H}}(\mathrm{d}h) \tag{3.29}$$

$$= -\int_H R(\mathbb{P}_\alpha^{\mathsf{D}|\mathsf{X}}, h)\mathbb{P}_C^{\mathsf{H}}(\mathrm{d}h) \qquad = -R_{\mathbb{P}_C^{\mathsf{H}}}(\mathbb{P}_\alpha^{\mathsf{D}|\mathsf{X}}) \tag{3.30}$$

□

Theorem 3.2.17 (Extension to a Bayesian see-do model)**.** *Given a see-do model* $(\mathbb{P}_{C\times H}, \mathsf{X}, \mathsf{Y}, \mathsf{D})$

Complete class theorem

A Bayesian see-do model is a probability set $\mathbb{P}_C$ and a see-do model is a probability set $\mathbb{P}_{C\times H}$. We can obtain the former by combining the latter with a prior $\mu \in \Delta(H)$; call this an induced Bayesian model with respect to μ. The *complete class theorem* establishes that any admissible decision rule for a see-do model $\mathbb{P}_{C\times H}$ must minimise the Bayes risk for the Bayesian model induced by some $\mu \in \Delta(H)$.

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 quotes are describing different things – in fact, they're not even describing the same *kind* of thing. The first is talking about 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 talking about a *function*, which is a purely mathematical object with a domain and a codomain and a mapping from the former into the latter.

The way we address this distinction is: a procedure that takes place in the real world and yields as results elements of a mathematical set is called a *measurement procedure*. We suppose for a given problem that there is a "complete measurement procedure" $\mathcal{S}$, and the result of every specific measurement procedure of interest can be reconstructed from the result of the complete measurement procedure by applying a function to the latter result. The function X that yields the result of a specific measurement procedure $\mathcal{X}$ given the result of the complete measurement procedure $\mathcal{S}$ is the *variable* associated with the measurement procedure $\mathcal{X}$.

In this way, the variable X – which is by itself just a mathematical function – is made relevant to the real-world by combining it with a total measurement procedure $\mathcal{S}$.

We can even use this scheme to address situations where it is possible to make different choices. We can simply posit a sub-procedure $\mathcal{C}$ that yields the choice α that we eventually make. However, modelling this can be tricky. If we want to use the consequence model to help make the decision, then it seems that the model of the decision procedure $\mathcal{C}$ will need to be self-referential. Furthermore, even if we have a model of $\mathcal{C}$ that says we will certainly decide on a particular element α^*, we still need to map every element of C to a consequence because this is what enables the comparison of elements of C. Thus, modelling $\mathcal{C}$ makes the model more complicated, and it's not obvious that this is answering the question that we are interested in. This complication is not obviously intractable, but we do not address it here. Instead, we simply assume that we have a collection $\mathcal{S}_\alpha$ of total measurement procedures that all yield elements of the same set, and each of which are executed if the choice made is α.

3.3.1 Variables and measurement procedures

We illustrate this approach with the example of Newton's second law in the form $\mathsf{F} = \mathsf{MA}$. This model relates "variables" F, M and A. As Feynman (1979) noted, in order to understand this law, we must bring some pre-existing understanding of force, mass and acceleration independent of the law itself. Furthermore, we contend, this knowledge cannot be expressed in any purely mathematical statement. In order to say what the net force on a given object is, even a highly knowledgeable physicist will have to go and do some measurements, which is a procedure that they carry out involving interacting with the real world somehow and obtaining as a result a vector representing the net forces on that object.

That is, the variables F, M and A are referring to the *results of measurement procedures*. We will introduce a separate notation to refer to these measurement procedures – $\mathcal{F}$ is the procedure for measuring force, $\mathcal{M}$ and $\mathcal{A}$ for mass and acceleration respectively. A measurement procedure $\mathcal{F}$ is akin to Menger (2003)'s notion of variables as "consistent classes of quantities" that consist of pairing between real-world objects and quantities of some type. Force $\mathcal{F}$ itself is not a well-defined mathematical thing, as measurement procedures are not mathematically well-defined. At the same time, the set of values it may yield *are* well-defined mathematical things. No actual procedure can be guaranteed to return elements of a mathematical set known in advance – anything can fail – but we assume that we can study procedures reliable enough that we don't lose much by making this assumption.

Note that, because $\mathcal{F}$ is not a purely mathematical thing, we cannot perform mathematical reasoning with $\mathcal{F}$ directly. Rather, we introduce a variable F which, as we will see, is a well-defined mathematical object, assert that it corresponds to $\mathcal{F}$ and conduct our reasoning using F.

3.3.2 Measurement procedures

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

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

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

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

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

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

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

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

For example, $\mathcal{MA}$ 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:

$$
\begin{aligned}
(g \circ f) \circ \mathcal{B} \text{ yields } x &\iff B \text{ yields } (g \circ f)^{-1}(x) && (3.31)\\
&\iff B \text{ yields } f^{-1}(g^{-1}(x)) && (3.32)\\
&\iff f \circ B \text{ yields } g^{-1}(x) && (3.33)\\
&\iff g \circ (f \circ B) \text{ yields } x && (3.34)
\end{aligned}
$$

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

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

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

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

3.3.3 Observable variables

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

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

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

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

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

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

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

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

3.3.4 Model variables

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

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

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

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

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

3.3.5 Variable sequences and partial order

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

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

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

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

3.3.6 Decision procedures

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

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

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

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

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

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

Chapter 4

Repeatable Response Functions

4.1 When do response functions exist?

We model decision problems with probability sets $\mathbb{P}_C$ for some set of choices C. If we have a pair of variables X and Y such that the uniform conditional $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ exists (Definition 2.4.3), then the joint outcome $\mathbb{P}_\alpha^{\mathsf{XY}}$ of any choice $\alpha \in C$ can be computed from the marginal distribution $\mathbb{P}_\alpha^{\mathsf{X}}$ alone.

We're interested in models that feature a particular kind of "causal effect" that we call a *conditionally independent and identical response functions*, or just "response functions" for short. They are a causal analogue of conditionally independent and identical sequences of random variables. Concretely, a model with response functions is a probability set $\mathbb{P}_C$ with variables $\mathsf{Y} := (\mathsf{Y}_i)_{i \in M}$, $(\mathsf{X}_i)_{i \in M}$ for some index set M and some H such that $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} C|\mathsf{X}_i\mathsf{H}$ and $\mathsf{H} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{X}_i C$ (see Section 2.4.2 for extended conditional independence) and $\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{H}} = \mathbb{P}_C^{\mathsf{Y}_j|\mathsf{X}_j\mathsf{H}}$ for all $i, j \in M$ (the identical response conditional requirement).

is that OK?

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

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

We prove our result under the simplifying assumption that $\mathsf{X}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}_{<i} C|\mathsf{X}_{<i}$. This is a limiting assumption – for example, it excludes cases where X_i depends

on $(\mathsf{X}_{<i}, \mathsf{Y}_{<i})$. In the more general case where this does not hold, the conditions we provide must still hold for any $C' \subset C$ such that $\mathsf{X}_i \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{Y}_{<i>} C | \mathsf{X}_{<i}$, but providing sufficient conditions in this case is the topic of further work.

Under this assumption, we show that *causal contractibility* is necessary and sufficient for the existence of response conditionals. Causal contractibility can be broken down into two sub-assumptions: *exchange commutativity* and *consequence locality*. The first is the assumption that the uniform conditional probability $\mathbb{P}_C^{\mathsf{Y}|\mathsf{X}}$ "commutes" with the permutation operation, and the second is the assumption that X_i "has no effect" on any Y_j for $j \neq i$.

4.1.1 Relevance to previous work

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

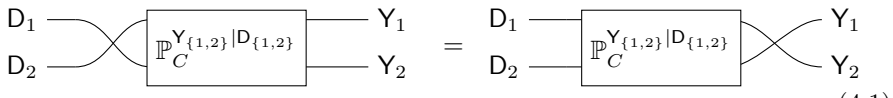
Post-treatment exchangeability found in Dawid (2020) is implied by exchange commutativity, but not the reverse. "Causal exchangeability" notions are also found in Greenland and Robins (1986) and Banerjee et al. (2017); a subtle difference between these notions and exchange commutativity is that these latter notions are given as symmetries of *decision procedures* – they involve actually swapping actions taken or individuals in an experiment – while exchange commutativity is a symmetry of probability sets.

Consequence locality is similar to the stable unit treatment distribution assumption (SUTDA) in Dawid (2020), although consequence locality is distinguished by being a concrete extended conditional independence (Definition 4.1.1) while SUTDA is given as the assumption that Y_i "depends only on" X_i. Consequence locality is also similar to stable unit treatment value assumption (SUTVA). The stable unit treatment value assumption (SUTVA) is given as (Rubin, 2005):

"("SUTVA) comprises two sub-assumptions. First, it assumes that *there is no interference between units (Cox 1958)*; that is, neither $Y_i(1)$ nor $Y_i(0)$ is affected by what action any other unit received. Second, it assumes that *there are no hidden versions of treatments*; no matter how unit i received treatment 1, the outcome that would be observed would be $Y_i(1)$ and similarly for treatment 0.

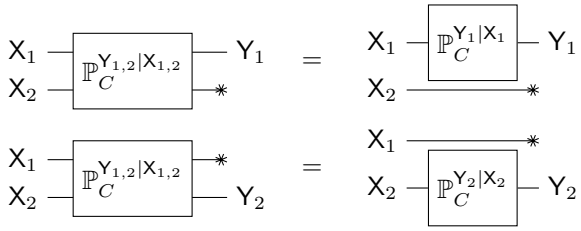
String diagram statements of both sub-assumptions can give some intuition about what they mean. For clarity, these diagrams illustrate the assumptions with exactly two inputs and two outputs, while the general definitions are for any countable number of inputs and outputs.

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



(4.1)

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



4.1.2 Causal contractibility

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

Definition 4.1.1 (Locality of consequences)**.** Suppose we have a sample space $(\Omega, \mathcal{F})$ and a probability set $\mathbb{P}_C$ where $\mathsf{Y} := \mathsf{Y} := (\mathsf{Y}_i)_M$, $\mathsf{D} := \mathsf{D}_M := (\mathsf{D}_i)_M$, $M \subseteq \mathbb{N}$. If for any $A \subset M$, $\mathsf{Y}_A \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{D}_{A^C} C | \mathsf{D}_A$ then $\mathbb{P}_C$ exhibits $(\mathsf{D};\mathsf{Y})$-*local consequences*.

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

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

Definition 4.1.2 (Swap map)**.** Given $M \subset \mathbb{N}$ a finite permutation $\rho : M \to M$ and a variable $\mathsf{X} : \Omega \to X^M$ such that $\mathsf{X} = (\mathsf{X}_i)_{i \in M}$, define the Markov kernel $\mathrm{swap}_{\rho(\mathsf{X})} : X^M \twoheadrightarrow X^M$ by $(d_i)_{i\in\mathbb{N}} \mapsto \delta_{(d_{\rho(i)})_{i\in\mathbb{N}}}$.

Definition 4.1.3 (Exchange commutativity)**.** Suppose we have a sample space $(\Omega, \mathcal{F})$ and a probability set $\mathbb{P}_C$ with uniform conditional probability $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ where $\mathsf{Y} := \mathsf{Y} := (\mathsf{Y}_i)_M$, $\mathsf{D} := \mathsf{D}_M := (\mathsf{D}_i)_M$, $M \subseteq \mathbb{N}$. If for any finite permutation $\rho : M \to M$

$$\mathrm{swap}_{\rho(\mathsf{D})} \mathbb{P}_C^{\mathsf{Y}|\mathsf{D}} \overset{\mathbb{P}_C}{\cong} \mathbb{P}_C^{\mathsf{Y}|\mathsf{D}} \mathrm{swap}_{\rho(\mathsf{Y})} \tag{4.4}$$

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

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

Theorem 4.1.4 shows that neither condition implies the other.

Theorem 4.1.4. *Exchange commutativity does not imply locality of consequences or vise versa.*

Proof. Appendix **??**. □

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

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

Although locality of consequences seems to intuitively encompass an assumption of non-interference, it still allows for some models in which exhibit certain kinds of interference between actions and outcomes of different indices. For example: I have an experiment where I first flip a coin and record the results of this flip as the outcome of the first step of the experiment, but I can choose either to record this same outcome as the provisional result of the second step (this is the choice $\mathsf{D}_1 = 0$), or choose to flip a second coin and record the result of that as the provisional result of the second step of the experiment (this is the choice $\mathsf{D}_1 = 1$). At the second step, I may further choose to copy the provisional results ($\mathsf{D}_2 = 0$) or invert them ($\mathsf{D}_2 = 1$). Then

$$\mathbb{P}_S^{\mathsf{Y}_1|\mathsf{D}}(y_1|d_1, d_2) = 0.5 \tag{4.5}$$

$$\mathbb{P}_S^{\mathsf{Y}_2|\mathsf{D}}(y_2|d_1, d_2) = 0.5 \tag{4.6}$$

- The marginal distribution of both experiments in isolation is Bernoulli(0.5) no matter what choices I make, so a model of this experiment would satisfies Definition 4.1.1
- Nevertheless, the choice for the first experiment affects the result of the second experiment

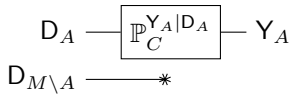
We call the conjunction of exchange commutativity and consequence locality *causal contractibility*.

Definition 4.1.5 (Causal contractibility)**.** A probability set $\mathbb{P}_C$ is $(\mathsf{D};\mathsf{Y})$-*causally contractible* if it is both exchange commutative and exhibits consequence locality.

Theorem 4.1.6 (Equality of reduced conditionals)**.** *A probability set $\mathbb{P}_C$ that is $(\mathsf{D};\mathsf{Y})$-causally contractible has, for any $A, B \subset M$ with $|A| = |B|$*

$$\mathbb{P}_C^{\mathsf{Y}_A|\mathsf{D}_A} \overset{\mathbb{P}_C}{\cong} \mathbb{P}_C^{\mathsf{Y}_B|\mathsf{D}_B} \tag{4.7}$$

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



$$= \quad \mathsf{D}_{M\setminus A}^{\mathsf{D}_A} \quad \mathbb{P}_C^{\mathsf{Y}_A\mathsf{Y}_{M\setminus A}|\mathsf{D}_A\mathsf{D}_{M\setminus A}} \quad \mathsf{Y}_A \tag{4.8}$$

$$= \quad \mathsf{D}_{M\setminus A}^{\mathsf{D}_A} \quad \mathrm{s}_{BA} \quad \mathbb{P}_C^{\mathsf{Y}_A\mathsf{Y}_{M\setminus A}|\mathsf{D}_A\mathsf{D}_{M\setminus A}} \quad \mathrm{s}_{AB} \quad \mathsf{Y}_A \tag{4.9}$$

$$= \quad \mathsf{D}_{M\setminus B}^{\mathsf{D}_B} \quad \mathbb{P}_C^{\mathsf{Y}_B\mathsf{Y}_{M\setminus B}|\mathsf{D}_A\mathsf{D}_{M\setminus B}} \quad \mathsf{Y}_B \tag{4.10}$$

Thus

$$\mathbb{P}_C^{\mathsf{Y}_A|\mathsf{D}_A\mathsf{D}_{M\setminus A}} \overset{\mathbb{P}_C}{\cong} \mathbb{P}_C^{\mathsf{Y}_B|\mathsf{D}_B\mathsf{D}_{M\setminus B}} \tag{4.11}$$

$$\overset{\mathbb{P}_C}{\cong} \quad \mathsf{D}_A \quad \mathbb{P}_C^{\mathsf{Y}_A|\mathsf{D}_A} \quad \mathsf{Y}_A \quad \mathsf{D}_{M\setminus A} \tag{4.12}$$

$$\implies \mathbb{P}_C^{\mathsf{Y}_A|\mathsf{D}_A} \overset{\mathbb{P}_C}{\cong} \mathbb{P}_C^{\mathsf{Y}_B|\mathsf{D}_B} \tag{4.13}$$

□

4.1.3 Existence of response conditionals

The main result in this section is Theorem 4.1.8 which shows that a probability set $\mathbb{P}_C$ is causally contractible if and only if it can be represented as the product of a distribution over hypotheses $\mathbb{P}_\square^{\mathsf{H}}$ and a collection of identical uniform conditionals $\mathbb{P}_C^{\mathsf{Y}_1|\mathsf{D}_1\mathsf{H}}$. Note the hypothesis H that appears in this conditional; it can be given the interpretation of a random variable that expresses the "true but initially unknown" $\mathsf{Y}_1|\mathsf{D}_1$ conditional probability.

Theorem 4.1.7. *Given a probability set $\mathbb{P}_C$ and variables $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$, $\mathbb{P}_C$ is $(\mathsf{D};\mathsf{Y})$-causally contractible if and only if there exists a column exchangeable probability distribution $\mu^{\mathsf{Y}^D} \in \Delta(Y^{|D|\times\mathbb{N}})$ such that*

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}} = \quad \mu^{\mathsf{Y}^D} \quad \mathsf{D} \quad \mathbb{F}_{\mathrm{ev}} \quad \mathsf{Y} \tag{4.14}$$

$$\iff \tag{4.15}$$

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}(y|(d_i)_{i\in\mathbb{N}}) = \mu^{\mathsf{Y}^D} \Pi_{(d_i i)_{i\in\mathbb{N}}}(y) \tag{4.16}$$

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

$$ev : D^{\mathbb{N}} \times Y^{D\times\mathbb{N}} \to Y \tag{4.17}$$
$$((d_i)_{\mathbb{N}}, (y_{ij})_{i\in D, j\in\mathbb{N}}) \mapsto (y_{d_i i})_{i\in\mathbb{N}} \tag{4.18}$$

Proof. Appendix **??**. □

We would prefer to talk about Y^D as a latent variable, rather than needing to refer to the the factorisation of a model in terms of μ^{Y^D} in Equation 4.14. This motivates the definition of an *augmented* causally contractible model.

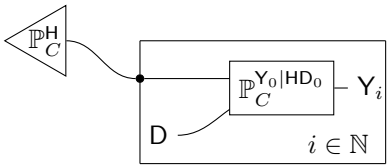
An augmented causally contractible model looks in some respects similar to a potential outcomes model - both have a distribution over an unobserved "tabular" variable Y^D, and the value of Y_i given D is deterministically equal to the $\mathsf{Y}_i^{\mathsf{D}}$ (abusing notation). However, the Y^D in an augmented causally contractible model usually can't be interpreted as potential outcomes. For example, consider a series of bets on fair coin flips. Model the consequence Y_i as uniform on $\{0,1\}$ for any decision D_i, for all i. Specifically, $D = Y = \{0,1\}$ and $\mathbb{P}_\alpha^{\mathsf{Y}_n}(y) = \prod_{i\in[n]} 0.5$ for all n, $y \in Y^n$, $\alpha \in R$. Then the construction of $\mathbb{P}^{\mathsf{Y}^D}$ following the method in Lemma 4.1.7 yields $\mathbb{P}^{Y_i{}^D}(y_i^D) = \prod_{j\in D} 0.5$ for all $y_i^D \in Y^D$. In this model Y_i^0 and Y_i^1 are independent and uniformly distributed. However, if we wanted Y_i^0 to be interpretable as "what would happen if I bet on outcome 0 on turn i" and Y^1 to represent "what would happen if I bet on outcome 1 on turn i", then we ought to have $\mathsf{Y}_i^0 = 1 - \mathsf{Y}_i^1$.

The following is the main theorem of this section, that establishes the equivalence between causal contractibility and the existence of response conditionals. The argument in outline is: because $\mathbb{P}_C^{\mathsf{Y}^D}$ is a column exchangeable probability distribution we can apply De Finetti's theorem to show $\mathbb{P}_C^{\mathsf{Y}^D}$ is representable as a product of identical parallel copies of $\mathbb{P}_C^{\mathsf{Y}_1^D|\mathsf{H}}$ and a common prior $\mathbb{P}_C^{\mathsf{H}}$. This in turn can be used to show that $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$ can be represented as a product of identical parallel copies of $\mathbb{P}_C^{\mathsf{Y}_1|\mathsf{D}_1\mathsf{H}}$ and the same common prior $\mathbb{P}_C^{\mathsf{H}}$.

Theorem 4.1.8. *Suppose we have a sample space* $(\Omega, \mathcal{F})$ *and a probability set* $\mathbb{P}_C$ *and variables* $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ *and* $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$. *Suppose also that* $\mathbb{P}_C$ *is* $(\mathsf{D};\mathsf{Y})$*-causally contractible if and only if there exists some* $\mathsf{H} : \Omega \to H$ *such that*

$\mathbb{P}_C^{\mathsf{H}}$ *and* $\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{HD}_i}$ *exist for all* $i \in \mathbb{N}$ *and*

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}} = \tag{4.19}$$



$$\iff \tag{4.20}$$

$$\mathsf{Y}_i \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{Y}_{\mathbb{N}\setminus i}, \mathsf{D}_{\mathbb{N}\setminus i} C | \mathsf{HD}_i \qquad \forall i \in \mathbb{N} \tag{4.21}$$

$$\wedge \mathsf{H} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{D}C \tag{4.22}$$

$$\wedge \mathsf{D}_i \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{Y}_{<i>} C | \mathsf{D}_{<i} \tag{4.23}$$

$$\wedge \mathbb{P}_C^{\mathsf{Y}_i|\mathsf{HD}_i} = \mathbb{P}^{\mathsf{Y}_0|\mathsf{HD}_0} \qquad \forall i \in \mathbb{N} \tag{4.24}$$

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

Proof. Appendix **??**. □

4.1.4 Elaborations and examples

Theorem 4.1.8 requires an infinite sequence of causally contractible pairs. In practice we only want to model finite sequences of variables, but this theorem applies as long as it is possible to extend the finite model to an infinite model maintaining causal contractibility.

Theorem 4.1.8 applies whatever procedure we use to obtain the $(\mathsf{D}_i, \mathsf{Y}_i)$ pairs – the D_is may be randomised, passive observations or active choices. Purely passive observations can be modeled with a probability set of size 1, and in this case an exchangeable sequence of $(\mathsf{D}_i, \mathsf{Y}_i)$ will also be causally contractible.

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

$$\mathbb{P}_\alpha^{\mathsf{D}_{[M+N]}} \text{swap}_\rho = \mathbb{P}_\alpha^{\mathsf{D}_{[M+N]}} \tag{4.25}$$

then by commutativity of exchange

$$\mathbb{P}_\alpha^{\mathsf{Y}_{[M+N]}} = \mathbb{P}_\alpha^{\mathsf{D}_{[M+N]}} \mathbb{P}_C^{\mathsf{Y}_{[M+N]}|\mathsf{D}_{[M+N]}} \tag{4.26}$$

$$= \mathbb{P}_\alpha^{\mathsf{D}_{[M+N]}} \text{swap}_\rho \mathbb{P}_C^{\mathsf{Y}_{[M+N]}|\mathsf{D}_{[M+N]}} \tag{4.27}$$

$$= \mathbb{P}_\alpha^{\mathsf{D}_{[M+N]}} \mathbb{P}_C^{\mathsf{Y}_{[M+N]}|\mathsf{D}_{[M+N]}} \text{swap}_\rho \tag{4.28}$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}_{[M+N]}} \text{swap}_\rho \tag{4.29}$$

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

$$\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}(A|d,h) = \int_X \mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{X}_i\mathsf{D}_i\mathsf{H}}(A|d,x,h)\mathbb{P}_\alpha^{\mathsf{X}_i|\mathsf{D}_i\mathsf{H}}(\mathrm{d}x|d,h) \tag{4.30}$$

$$= \int_X \mathbb{P}_C^{\mathsf{Y}_1|\mathsf{X}_1\mathsf{D}_1\mathsf{H}}(A|d,x,h)\mathbb{P}_C^{\mathsf{X}_i|\mathsf{H}}(\mathrm{d}x|h) \tag{4.31}$$

If we additionally assume $\mathbb{P}_C^{\mathsf{X}_i|\mathsf{H}} \cong \mathbb{P}_C^{\mathsf{X}_1|\mathsf{H}}$ then

$$\mathbb{P}_\alpha^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}(A|d,h) = \int_X \mathbb{P}_C^{\mathsf{Y}_1|\mathsf{X}_1\mathsf{D}_1\mathsf{H}}(A|d,x,h)\mathbb{P}_C^{\mathsf{X}_1|\mathsf{H}}(\mathrm{d}x|h) \tag{4.32}$$

Equation 4.32 is identical to the backdoor adjustment formula for an intervention on D_1 targeting Y_1 where X_1 is a common cause of both.

While it is formally possible to use a causally contractible model for a decision procedure that involves both passive observations and active choices, causal contractibility is a very strong assumption. Suppose we have a decision procedure in which M passive observations are made $(\mathsf{D}_M, \mathsf{Y}_M)$, followed by M active choices $(\mathsf{D}_{(M,2M]}, \mathsf{Y}_{(M,2M]})$. If a model $\mathbb{P}_C$ of this procedure is $(\mathsf{D}_{2M}, \mathsf{Y}_{2M})$-causally contractible model then the following holds (see corollary 4.3.5):

$$\mathbb{P}_C^{\mathsf{Y}_{[2,M+1]}|\mathsf{D}_{[2,M+1]}} = \mathbb{P}^{\mathsf{Y}_{(M,2M]}|\mathsf{D}_{(M,2M]}} \tag{4.33}$$

$$\implies \mathbb{P}_C^{\mathsf{Y}_{M+1}|\mathsf{D}_{[2,M+1]}\mathsf{Y}_{[2,M]}} = \mathbb{P}^{\mathsf{Y}_{M+1}|\mathsf{D}_{(M,2M]}\mathsf{Y}_{(M+1,2M]}} \tag{4.34}$$

That is, causal contractibility implies that there is no difference between conditioning on observational results or on the results of active choices; active choices are as good for predicting observations as vise-versa. Normally one might consider randomised experimental results to be "better" than passive observations, but this is not compatible with the assumption of causal contractibility.

4.1.5 Assessing causal contractibility

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

Two step justifications of this form are common in literature on causal identifiability. For example, Greenland and Robins (1986) explain

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

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

Dawid (2020) suggests (with many qualifications) that "post-treatment exchangeability" for a decision problem regarding taking aspirin to treat a headache may be acceptable if the data are from

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

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

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

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

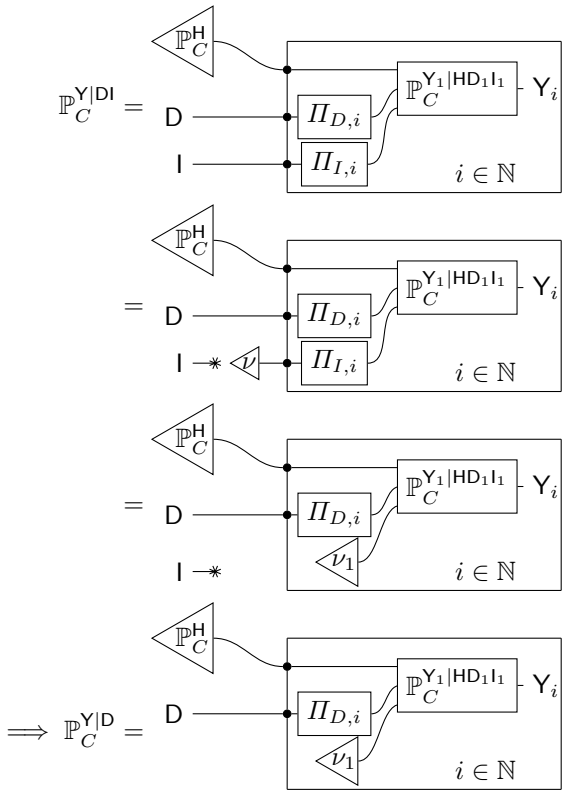
Here we have a more abstract statement about the row-exchangeability of "the science", rather than individual people involved in an experiment, but we regard it as similar in spirit to assumptions that people involved in the experiment are "similar". Rubin explicitly mentions a second condition: that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

Theorem 4.1.12 formalises these ideas. As an example of its application, consider an experiment where N patients, each with an individual identifier I_i, receive treatment D_i and experience outcome Y_i. We assume a $((\mathsf{D}, \mathsf{I}); \mathsf{Y})$-causally contractible model $\mathbb{P}_C$ is appropriate. This reflects two judgments; firstly, that treatment D_i and identifiers I_i screen off all other variables from Y_i (Definition 4.1.1), and secondly that the order in which the individuals appear and the treatments are received does not alter the consequences we expect to see (Definition 4.1.3). The fact that we need a preliminary assumption of causal contractibility is similar to how, in the potential outcomes framework, a preliminary assumption of SUTVA is required in order to justify the use of potential outcomes.

Next, we assume that, no matter which choice $\alpha \in C$ is decided on, all identifiers can be swapped without altering the distribution over consequences, and finally that for each choice $\alpha \in C$ the treatment vector D is deterministic. Then, according to Theorem 4.1.12, $\mathbb{P}_C$ is also $(\mathsf{D}; \mathsf{Y})$-causally contractible. This can be extended to the case where D is a function of a "random signal" R.

Lemma 4.1.9. *If $\mathbb{P}_C$ is $((\mathsf{D},\mathsf{I});\mathsf{Y})$-causally contractible and $\mathsf{Y} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{I}C|\mathsf{D}$, then $\mathbb{P}_C$ is also $(\mathsf{D};\mathsf{Y})$-causally contractible.*

Proof. For arbitrary $\nu \in \Delta(I^{\mathbb{N}})$, by assumption of $((\mathsf{D},\mathsf{I});\mathsf{Y})$-causal contractibility and Theorem 4.1.8



Applying Theorem 4.1.8 in reverse, we get $\mathbb{P}_C$ is $(\mathsf{D};\mathsf{Y})$-causally contractible. □

Definition 4.1.10 (Index variable)**.** Suppose we have a probability set $\mathbb{P}_C$ and a variable $\mathsf{I} : \Omega \to \mathbb{N}^{\mathbb{N}}$, such that, defining $A \subset \mathcal{P}(\mathbb{N})$ to be the set of all permutations of $\mathbb{N}$, $\mathbb{P}^{\mathsf{I}}_{\alpha}(A) = 1$. Then I is an *index variable.*

Lemma 4.1.11. *Suppose we have a probability set $\mathbb{P}_C$ where $\mathsf{Y} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} C|(\mathsf{D},\mathsf{I})$ and I is an index variable. If for each permutation $\rho : \mathbb{N} \to \mathbb{N}$*

$$\mathbb{P}^{\mathsf{Y}|\mathsf{ID}}_{\alpha} = (swap_{\rho(I)} \otimes Id_X)\mathbb{P}^{\mathsf{Y}|\mathsf{ID}}_{\alpha} \tag{4.39}$$

then $\mathsf{Y} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{I}C|\mathsf{D}$.

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

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

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{ID}}(B|i,d) = (\mathrm{swap}_{\rho_i(I)} \otimes \mathrm{Id}_X)\mathbb{P}_C^{\mathsf{Y}|\mathsf{ID}}(B|i,d) \tag{4.40}$$

$$= \mathbb{P}_C^{\mathsf{Y}|\mathsf{ID}}(B|\mathrm{id}_{\mathbb{N}},d) \tag{4.41}$$

Therefore

$$\mathbb{P}_C^{\mathsf{Y}|\mathsf{ID}} \overset{\mathbb{P}_C}{\cong} \mathrm{erase}_{\mathbb{N}^{\mathbb{N}}} \otimes \mathbb{K} \tag{4.42}$$

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

$$(B|d) \mapsto \mathbb{P}_C^{\mathsf{Y}|\mathsf{ID}}(B|\mathrm{id}_{\mathbb{N}},d) \tag{4.43}$$

□

Theorem 4.1.12. *Suppose we have a probability set $\mathbb{P}_C$, C countable, that $((\mathsf{D},\mathsf{I});\mathsf{Y})$-causally contractible for variables $\mathsf{Y} : \Omega \to Y^{\mathbb{N}}$, $D : \Omega \to D^{\mathbb{N}}$ and index variable $\mathsf{I} : \Omega \to \mathbb{N}^{\mathbb{N}}$. If for each $\alpha \in C$, $\rho : \mathbb{N} \to \mathbb{N}$*

$$\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{I}} = swap_{\rho(I)}\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{I}} \tag{4.44}$$

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

$$\mathbb{P}_\alpha^{\mathsf{D}} = \mathbb{F}_f \tag{4.45}$$

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

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

$$\mathbb{Q} \overset{\mathbb{P}_\alpha}{\cong} (\mathrm{id}_{\mathbb{N}^{\mathbb{N}}} \otimes \mathbb{F}_f)\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{ID}} \tag{4.46}$$

Furthermore, by assumption

$$(\mathrm{swap}_{\rho(I)} \otimes \mathrm{Id}_C)\mathbb{Q} = \mathbb{Q} \tag{4.47}$$

Therefore

$$(\mathrm{swap}_{\rho(I)} \otimes \mathrm{Id}_C)\mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{ID}} = (\mathrm{id}_{\mathbb{N}^{\mathbb{N}}} \otimes \mathbb{F}_f)(\mathrm{id}_{\mathbb{N}^{\mathbb{N}}} \otimes \mathbb{F}_{f^{-1}})(\mathrm{swap}_{\rho(I)} \otimes \mathrm{Id}_C)\mathbb{Q} \tag{4.48}$$

$$= (\mathrm{id}_{\mathbb{N}^{\mathbb{N}}} \otimes \mathbb{F}_{f^{-1}})\mathbb{Q} \tag{4.49}$$

$$= \mathbb{P}_\alpha^{\mathsf{Y}|\mathsf{ID}} \tag{4.50}$$

Then by Lemma 4.1.11 we have $\mathsf{Y} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{I}C|\mathsf{D}$ and by Lemma 4.1.9 we have $(\mathsf{D};\mathsf{Y})$-causal contractibility. □

If we suppose Theorem 4.1.12 applies to $C' \subset C$ such that D is deterministic for all $\alpha \in C'$, while C consists of C' and all "random choices between elements of C'"; that is for all $\beta \in C$

$$\mathbb{P}_\beta = \sum_{c \in C} k_c \mathbb{P}_c \tag{4.51}$$

Then it follows that

$$\mathbb{P}_\beta^{\mathsf{Y}|\mathsf{D}} = \sum_{c \in C} k_c \mathbb{P}_c^{\mathsf{Y}|\mathsf{D}} \tag{4.52}$$

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

4.1.6 Body mass index revisited

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

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

Suppose we are in possession of a $(\mathsf{D};(\mathsf{B},\mathsf{Y}))$-causally contractible probability set $\mathbb{P}_C$, such that each $(\mathsf{D}_i, \mathsf{B}_i, \mathsf{Y}_i)$ is related by the response conditional $\mathbb{P}_C^{\mathsf{Y}_1\mathsf{B}_1|\mathsf{D}_1\mathsf{H}}$. Suppose that we also have an "oracle" available that performs an infinite number of samples under appropriate conditions and reveals the value $h \in H$ yielded by the variable H. Then we can consider the new probability set $\mathbb{P}_{C,h}$ where for arbitrary $\mathsf{Z} : \Omega \to Z$, $A \in \mathcal{Z}$

$$\mathbb{P}_{C,h}^{\mathsf{Z}}(A) = \mathbb{P}_C^{\mathsf{Z}|\mathsf{H}}(A|h) \tag{4.53}$$

Note that $\mathbb{P}_{C,h}$ remains $(\mathsf{D};(\mathsf{B},\mathsf{Y}))$-causally contractible with response conditionals $\mathbb{P}_{C,h}^{\mathsf{Y}_1\mathsf{B}_1|\mathsf{D}_1}$. Furthermore that by Theorem 2.4.27, we have $((\mathsf{D},\mathsf{B});\mathsf{Y})$-causal contractibility with response conditionals $\mathbb{P}_{C,h}^{\mathsf{Y}_1|\mathsf{D}_1\mathsf{B}_1}$. In this case, we could find that $\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_{C,h}} \mathsf{D}_i|\mathsf{B}_i$, and so by Lemma 4.1.9 $\mathbb{P}_{C,h}$ is also $(\mathsf{B};\mathsf{Y})$-causally contractible.

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

4.2 Allowing dependence on observations

We want to remove the assumption $\mathsf{D}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}_{<i} C | \mathsf{D}_{<i}$. It turns out that this is fairly straightforward to do – essentially, we require that causal contractibility holds for some C' such that $\mathsf{D}_i \perp\!\!\!\perp^e_{\mathbb{P}_{C'}} \mathsf{Y}_{<i} C | \mathsf{D}_{<i}$, and then we require that $\mathbb{P}_C$ extends $\mathbb{P}_{C'}$ in the appropriate way. Before we do this, however, we will introduce the idea of a *comb*, which is an important generalisation of the idea of a conditional probability. The result we will eventually arrive at is that a model features response conditionals if it has a causally contractible comb of the appropriate type.

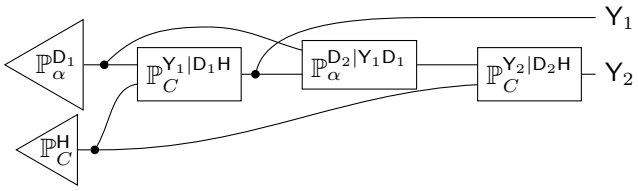
To begin with an example, consider a probability set $\mathbb{P}_C$, variables $\mathsf{D}_\mathbb{N}$ and $\mathsf{Y}_\mathbb{N}$ and a subsequence $(\mathsf{D}_i, \mathsf{Y}_i)_{i \in [2]}$ of length 2. Lifting the assumption $\mathsf{D}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}_{<i} C | \mathsf{D}_{<i}$ means that only the following holds:

$$\mathsf{Y}_i \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{Y}_{[\mathbb{N}]\setminus i}, \mathsf{D}_{\mathbb{N}\setminus i} C | \mathsf{H}\mathsf{D}_i \qquad \forall i \in \mathbb{N} \tag{4.54}$$

$$\wedge \mathsf{H} \perp\!\!\!\perp^e_{\mathbb{P}_C} \mathsf{D} C \tag{4.55}$$

$$\wedge \mathbb{P}_C^{\mathsf{Y}_i | \mathsf{H}\mathsf{D}_i} = \mathbb{P}^{\mathsf{Y}_0 | \mathsf{H}\mathsf{D}_0} \qquad \forall i \in \mathbb{N} \tag{4.56}$$

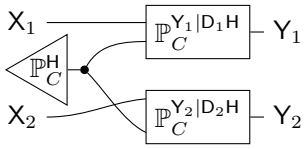
In this case, for arbitrary $\alpha \in C$

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

(4.57)

we still have response conditionals $\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}$ but now the D_i's can depend on the previous Y_i's.

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

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

(4.58)

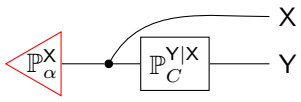
and $\mathbb{P}_C^{\mathsf{Y}_{[2]} \wr \mathsf{D}_{[2]}}$ is still "$(\mathsf{D};\mathsf{Y})$-causally contractible", but it is no longer a uniform conditional probability. It is is a *uniform 2-comb*.

4.2.1 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. Similarly, higher

order combs map Graphically, the semidirect product looks like

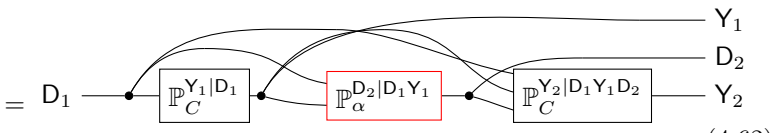
$$\mathbb{P}_\alpha^{\mathsf{XY}} = \mathbb{P}_\alpha^{\mathsf{X}} \odot \mathbb{P}_C^{\mathsf{Y}|\mathsf{X}} \tag{4.59}$$



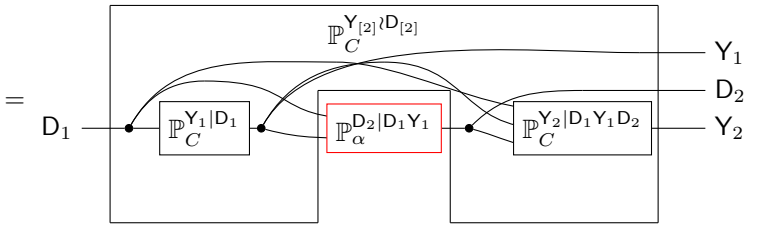
$$= \tag{4.60}$$

and the insert operation looks like

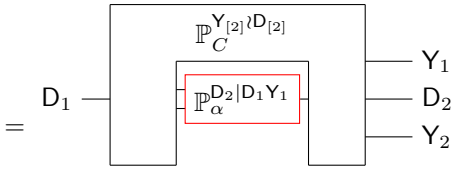
$$\mathbb{P}_\alpha^{\mathsf{Y}_1\mathsf{D}_2\mathsf{Y}_2|\mathsf{D}_1} = \text{insert}(\mathbb{P}_\alpha^{\mathsf{D}_2|\mathsf{D}_1\mathsf{Y}_1}, \mathbb{P}_C^{\mathsf{Y}_{[2]}\wr\mathsf{D}_{[2]}}) \tag{4.61}$$



(4.62)



(4.63)



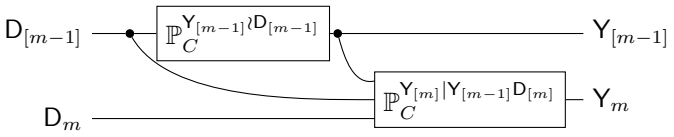
(4.64)

While Equation 4.62 is a well-formed string diagram in the category of Markov kernels, Equation 4.64 is not. In the case that all the underlying sets are discrete, Equation 4.64 can be defined using an extended string diagram notation appropriate for the category of real-valued matrices (Jacobs et al., 2019). We're not going to go to introduce the extension here, but we will give an algebraic definition of the insert operation for discrete sets.

Definition 4.2.1 (Uniform n-Comb)**.** Given a probability set $\mathbb{P}_C$ with variables $\mathsf{Y}_i : \Omega \to Y$ and $\mathsf{D}_i : \Omega \to D$ for $i \in [n]$ an uniform conditional probabilities $\{\mathsf{P}_C^{\mathsf{Y}_i|\mathsf{D}_{[i]}\mathsf{Y}_{[i-1]}} | i \in [n]\}$, the uniform n-comb $\mathbb{P}_C^{\mathsf{Y}_{[n]}\wr\mathsf{D}_{[n]}} : D^n \twoheadrightarrow Y^n$ is the Markov

kernel given by the recursive definition

$$\mathbb{P}_C^{\mathsf{Y}_1 \wr \mathsf{D}_1} = \mathbb{P}_C^{\mathsf{Y}_1 | \mathsf{D}_1} \tag{4.65}$$



$$\mathbb{P}_C^{\mathsf{Y}_{[m]} \wr \mathsf{D}_{[m]}} = \tag{4.66}$$

Definition 4.2.2 (Uniform $\mathbb{N}$-comb)**.** Given a probability set $\mathbb{P}_C$ with variables $\mathsf{Y}_i : \Omega \to Y$ and $\mathsf{D}_i : \Omega \to D$ for $i \in \mathbb{N}$ and uniform conditional probabilities $\{\mathsf{P}_C^{\mathsf{Y}_i | \mathsf{D}_{[i]} \mathsf{Y}_{[i-1]}} | i \in \mathbb{N}\}$, the uniform $\mathbb{N}$-comb $\mathbb{P}_C^{\mathsf{Y}_{\mathbb{N}} \wr \mathsf{D}_{\mathbb{N}}} : D^{\mathbb{N}} \nrightarrow Y^{\mathbb{N}}$ is the Markov kernel such that for all $n \in \mathbb{N}$

$$\mathbb{P}_C^{\mathsf{Y}_{\mathbb{N}} \wr \mathsf{D}_{\mathbb{N}}}(\mathrm{id}_{Y^n} \otimes \mathrm{del}_{Y^{\mathbb{N}}}) = \mathbb{P}_C^{\mathsf{Y}_{[n]} \wr \mathsf{D}_{[n]}} \otimes \mathrm{del}_{Y^{\mathbb{N}}} \tag{4.67}$$

Theorem 4.2.3 (Existence of $\mathbb{N}$-combs)**.** *Given a probability set $\mathbb{P}_C$ with variables $\mathsf{Y}_i : \Omega \to Y$ and $\mathsf{D}_i : \Omega \to D$ for $i \in \mathbb{N}$ and uniform conditional probabilities $\{\mathsf{P}_C^{\mathsf{Y}_i | \mathsf{D}_{[i]} \mathsf{Y}_{[i-1]}} | i \in \mathbb{N}\}$, a uniform $\mathbb{N}$-comb $\mathbb{P}_C^{\mathsf{Y}_{\mathbb{N}} \wr \mathsf{D}_{\mathbb{N}}} : D^{\mathbb{N}} \nrightarrow Y^{\mathbb{N}}$ exists.*

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

$$\mathbb{P}_C^{\mathsf{Y}_{[n]} \wr \mathsf{D}_{[n]}}(\mathrm{id}_{Y^{n-m}}) \otimes \mathrm{del}_{Y^m}) = \mathbb{P}_C^{\mathsf{Y}_{[n-m]} \wr \mathsf{D}_{[n-m]}} \otimes \mathrm{del}_{Y^m} \tag{4.68}$$

Therefore the existence of $\mathbb{P}_C^{\mathsf{Y}_{\mathbb{N}} \wr \mathsf{D}_{\mathbb{N}}}$ is a consequence of Lemma 4.3.8. □

We now define the insert operation for discrete sets only

Definition 4.2.4 (Comb insert)**.** Given an n-comb $\mathbb{P}_C^{\mathsf{Y}_{[n]} \wr \mathsf{D}_{[n]}}$ and an $n-1$ comb $\mathbb{P}_\alpha^{\mathsf{D}_{[1,n]} | \mathsf{Y}_{[n-1]}}$, $(D, \mathcal{D})$ and $(Y, \mathcal{Y})$ discrete, for all $y_i \in Y$ and $d_i \in D$

$$\mathrm{insert}(\mathbb{P}_\alpha^{\mathsf{D}_{[1,n]} \wr \mathsf{Y}_{[n-1]}}, \mathbb{P}_C^{\mathsf{Y}_{[n]} \wr \mathsf{D}_{[n]}})(y_{[n]}, d_{[n-1]} | d_1) = \mathbb{P}_C^{\mathsf{Y}_{[n]} \wr \mathsf{D}_{[n]}}(y_n | d_{[1,n]}, d_1) \mathbb{P}_\alpha^{\mathsf{D}_{[1,n]} \wr \mathsf{Y}_{[n-1]}}(d_{[1,n]} | y_{[n-1]}) \tag{4.69}$$

4.2.2 Response conditionals in models with history dependence

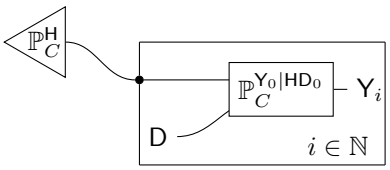
The main theorem of this section, Theorem 4.2.5 is an almost trivial extension of Theorem 4.1.8. Instead of starting with a probability set $\mathbb{P}_C$ with a uniform conditional $\mathbb{P}_C^{\mathsf{Y} | \mathsf{D}}$, we assume instead a collection of uniform conditionals $\mathbb{P}_C^{\mathsf{Y}_i | \mathsf{D}_{\leq i} \mathsf{Y}_{<i}}$ (this is a strictly weaker condition, as we can derive these as higher order conditionals from the assumption of $\mathbb{P}_C^{\mathsf{Y} | \mathsf{D}}$). Theorem 4.2.5 then says if the $\mathbb{N}$-comb $\mathbb{P}_C^{\mathsf{Y} \wr \mathsf{D}}$ is causally contractible, then we have response conditionals

$\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{D}_i\mathsf{H}}$. Because $\mathbb{P}_C^{\mathsf{Y}\wr\mathsf{D}}$ is a Markov kernel with the same signature as $\mathbb{P}_C^{\mathsf{Y}|\mathsf{D}}$, the reasoning from Theorem 4.2.5 carries over almost directly..

Unlike causal contractibility in the history independent case, we are not aware of an equivalent condition to the causal contractibility of $\mathbb{P}_C^{\mathsf{Y}\wr\mathsf{D}}$ in terms of relabeled random variables. The difficulty is, because D_i can depend on $\mathsf{Y}_{<i}$, we usually don't have symmetry with respect to relabeling variables.

Theorem 4.2.5. *Suppose we have a sample space $(\Omega, \mathcal{F})$, a probability set $\mathbb{P}_C$ and variables $\mathsf{D} := (\mathsf{D}_i)_{i\in\mathbb{N}}$ and $\mathsf{Y} := (\mathsf{Y}_i)_{i\in\mathbb{N}}$ with uniform $\mathbb{N}$-comb $\mathbb{P}_C^{\mathsf{Y}\wr\mathsf{D}}$. $\mathbb{P}_C^{\mathsf{Y}\wr\mathsf{D}}$ is causally contractible if and only if Ω can be extended with some $\mathsf{H} : \Omega \times H \to H$ such that $\mathbb{P}_C^{\mathsf{H}}$ and $\mathbb{P}_C^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i}$ exist for all $i \in \mathbb{N}$ and*

$$\mathbb{P}_C^{\mathsf{Y}\wr\mathsf{D}} = \tag{4.70}$$



$$\iff \tag{4.71}$$

$$\mathsf{Y}_i \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{Y}_{\mathbb{N}\setminus i}, \mathsf{D}_{\mathbb{N}\setminus i} C | \mathsf{H}\mathsf{D}_i \qquad \forall i \in \mathbb{N} \tag{4.72}$$

$$\wedge \mathsf{H} \perp\!\!\!\perp^{e}_{\mathbb{P}_C} \mathsf{D}C \tag{4.73}$$

$$\wedge \mathbb{P}_C^{\mathsf{Y}_i|\mathsf{H}\mathsf{D}_i} = \mathbb{P}^{\mathsf{Y}_0|\mathsf{H}\mathsf{D}_0} \qquad \forall i \in \mathbb{N} \tag{4.74}$$

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

lemma I've yet to prove

Proof. Apply Lemma to $\mathbb{P}_C^{\mathsf{Y}\wr\mathsf{D}}$. □

4.2.3 Validity

4.2.4 Combs are the output of the "fix" operation

There is a relationship between combs and the "fix" operation defined in Richardson et al. (2017). In particular, suppose we have a probability $\mathbb{P}_\alpha$ and a comb $\mathbb{P}_\alpha^{\mathsf{Y}_{[2]}|\mathsf{D}_{[2]}}$. Then (assuming discrete sets)

$$\mathbb{P}_\alpha^{\mathsf{Y}_{[2]}|\mathsf{D}_{[2]}}(y_1, y_2|d_1, d_2) = \mathbb{P}_\alpha^{\mathsf{Y}_1|\mathsf{D}_1}(y_1|d_1)\mathbb{P}_\alpha^{\mathsf{Y}_2|\mathsf{D}_2}(y_2|d_2) \tag{4.75}$$

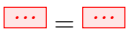
$$= \frac{\mathbb{P}_\alpha^{\mathsf{Y}_1|\mathsf{D}_1}(y_1|d_1)\mathbb{P}_\alpha^{\mathsf{D}_2|\mathsf{Y}_1\mathsf{D}_1}(d_2|y_1, d_1)\mathbb{P}_\alpha^{\mathsf{Y}_2|\mathsf{D}_2}(y_2|d_2)}{\mathbb{P}_\alpha^{\mathsf{D}_2|\mathsf{Y}_1\mathsf{D}_1}(d_2|y_1, d_1)} \tag{4.76}$$

$$= \frac{\mathbb{P}_\alpha^{\mathsf{Y}_{[}2]\mathsf{D}_2|\mathsf{D}_1}(y_1, y_2, d_2|d_1)}{\mathbb{P}_\alpha^{\mathsf{D}_2|\mathsf{Y}_1\mathsf{D}_1}(d_2|y_1, d_1)} \tag{4.77}$$

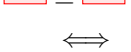
This is precisely the "division by a conditional probability" used in the fix operation. We speculate that the fix operation is precisely an alternative definition of an n-comb, but we have not proven this.

4.3 Weaker assumptions than causal contractibility

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



$$= \tag{4.78}$$



$$\iff \tag{4.79}$$

$$\mathbb{K}(\bigtimes_{i\in[n]} A_i \times Y^{\mathbb{N}} | x_{[n]}, x_{[n]^C}) = \mathbb{L}(\bigtimes_{i\in[n]} A_i | x_{[n]}) \tag{4.80}$$

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

$$\mathbb{K}\mathrm{swap}_{\rho_Y} = \mathrm{swap}_{\rho_X}\mathbb{K} \tag{4.81}$$

$$\iff \tag{4.82}$$

$$\mathbb{K}(\bigtimes_{i\in\mathbb{N}} A_{\rho(i)} | \bigotimes_{i\in\mathbb{N}} x_i) = \mathbb{K}(\bigtimes_{i\in\mathbb{N}} A_i | \bigotimes_{i\in\mathbb{N}} x_{\rho(i)}) \tag{4.83}$$

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

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

Theorem 4.3.5 (Equality of equally sized contractions)**.** *A Markov kernel* $\mathbb{K} : X^{\mathbb{N}} \nrightarrow Y^{\mathbb{N}}$ *is* causally contractible *if and only if for every* $n \in \mathbb{N}$ *there exists there exists* $\mathbb{L}_n : X^n \nrightarrow Y^n$ *such that for all* $A \subset \mathbb{N}$ *where* $|A| = n$

$$\mathit{swap}_{[n]\leftrightarrow A}\mathbb{K}\mathit{swap}_{[n]\leftrightarrow A}(\mathit{id}_{[n]} \otimes \mathit{del}_{[n]^C}) = \mathbb{L}_n \otimes \mathrm{del}_{[n]^C} \tag{4.84}$$

Proof. Only if: By exchange commutativity

$$\mathrm{swap}_{[n]\leftrightarrow A}\mathbb{K} = \mathbb{K}\mathrm{swap}_{A\leftrightarrow n} \tag{4.85}$$

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

$$\mathrm{swap}_{[n]\leftrightarrow A}\mathbb{K}\mathrm{swap}_{A\leftrightarrow n} = \mathbb{K} \tag{4.86}$$

Then by locality, there exists some $\mathbb{L} : X^n \nrightarrow Y^n$ such that

$$\mathrm{swap}_{[n]\leftrightarrow A}\mathbb{K}\mathrm{swap}_{A\leftrightarrow n}(\mathrm{id}_{[n]} \otimes \mathrm{del}_{[n]^C}) = \mathbb{K}(\mathrm{id}_{[n]} \otimes \mathrm{del}_{[n]^C}) \tag{4.87}$$

$$= \mathbb{L} \otimes \mathrm{del}_{[n]^C} \tag{4.88}$$

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

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

$$\mathrm{swap}_{A \leftrightarrow [n]} \mathbb{K} \mathrm{swap}_{A \leftrightarrow [n]} (\mathrm{id}_{[n]} \otimes \mathrm{del}_{[n]^C}) = \mathbb{K}(\mathrm{id}_{[n]} \otimes \mathrm{del}_{[n]^C}) \tag{4.89}$$

Then by Lemma 4.3.8

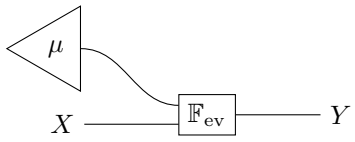
$$\mathrm{swap}_{A \leftrightarrow [n]} \mathbb{K} \mathrm{swap}_{A \leftrightarrow [n]} = \mathbb{K} \tag{4.90}$$

Consider an arbitrary finite permutation $\rho : \mathbb{N} \to \mathbb{N}$. ρ can be decomposed into a finite set of cyclic permutations on disjoint orbits. Each cyclic permutation is simply the composition of a sequence of 1-cycles of the form $A \leftrightarrow [n]$, and so ρ itself can be written as a composition of a sequence of 1-cycles. Thus for any finite $\rho : \mathbb{N} \to \mathbb{N}$

$$\mathrm{swap}_{\rho} \mathbb{K} \mathrm{swap}_{\rho} = \mathbb{K} \tag{4.91}$$

□

Theorem 4.3.6. *A Markov kernel $\mathbb{K} : X^{\mathbb{N}} \nrightarrow Y^{\mathbb{N}}$ is causally contractible if and only if there exists a column exchangeable probability distribution $\mu \Delta(Y^{|X| \times \mathbb{N}})$ such that*

$\mathbb{K} =$ 

(4.92)

$$\iff \tag{4.93}$$

$$\mathbb{K}(A|(x_i)_{i \in \mathbb{N}}) = \mu \Pi_{(x_i i)_{i \in \mathbb{N}}}(A) \forall A \in \mathcal{Y}^{\mathbb{N}} \tag{4.94}$$

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

$$(y_{ji})_{j,i \in X \times \mathbb{N}} \mapsto (y_{d_i i})_{i \in \mathbb{N}} \tag{4.95}$$

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

$$ev : X^{\mathbb{N}} \times Y^{X \times \mathbb{N}} \to Y \tag{4.96}$$

$$((x_i)_{\mathbb{N}}, (y_{ji})_{j,i \in X \times \mathbb{N}}) \mapsto (y_{x_i i})_{i \in \mathbb{N}} \tag{4.97}$$

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

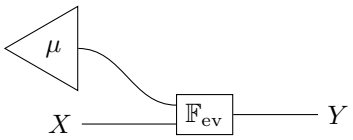
Define

$$\mu(\bigtimes_{(i,j) \in X \times \mathbb{N}} A_{ij}) := \mathbb{K}(\bigtimes_{(i,j) \in X \times \mathbb{N}} A_{ij} | e) \forall A_{ij} \in \mathcal{Y} \tag{4.98}$$

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

$$\mathbb{Q} : X^{\mathbb{N}} \twoheadrightarrow Y^{\mathbb{N}} \tag{4.99}$$

$\mathbb{Q} :=$ 

(4.100)

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

$$\mathbb{Q}(C_A | x) = \mu(\text{ev}_x^{-1}(C_A)) \tag{4.101}$$

The argument of μ is

$$\begin{aligned} \text{ev}_x^{-1}(C_A) &= \{(y_{ji})_{j,i \in X \times \mathbb{N}} | (y_{x_i i})_{i \in \mathbb{N}} \in C_A\} && (4.102)\\ &= \bigtimes_{i \in \mathbb{N}} \bigtimes_{j \in X} D_{ji} && (4.103) \end{aligned}$$

where

$$D_{ji} = \begin{cases} C_i & (j,i) \in B \\ Y & \text{otherwise} \end{cases} \tag{4.104}$$

and so

$$\text{swap}_{A \leftrightarrow B}(\text{ev}_x^{-1}(C_A)) = C_A \tag{4.105}$$

Substituting Equation 4.105 into 4.101

$$\begin{aligned} \mathbb{Q}(C_A | x) &= \mu \text{swap}_{A \leftrightarrow B}(C_A) && (4.106)\\ &= \mathbb{K} \text{swap}_{A \leftrightarrow B}(C_A | e) && (4.107)\\ &= \mathbb{K} \text{swap}_{A \leftrightarrow B}(C_A | e_B, \text{swap}_{B \leftrightarrow A}(x)_B^C) && \text{by locality} \\ & && (4.108)\\ &= \mathbb{K} \text{swap}_{A \leftrightarrow B}(C_A | \text{swap}_{B \leftrightarrow A}(x)) && (4.109)\\ &= \text{swap}_{B \leftrightarrow A} \mathbb{K} \text{swap}_{A \leftrightarrow B}(C_A | x) && (4.110)\\ &= \mathbb{K}(C_A | x) && \text{by commutativity of exchange} \\ & && (4.111) \end{aligned}$$

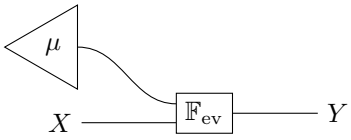
Because this holds for all x, $A \subset \mathbb{N}$, by Lemma 4.3.8

$$\mathbb{Q} = \mathbb{K} \tag{4.112}$$

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

$$\mathbb{K}(A|e) = \text{swap}_c \mathbb{K} \text{swap}_c(A|e) \tag{4.113}$$
$$= \mathbb{K} \text{swap}_c(A|\text{swap}_c(e)) \tag{4.114}$$
$$= \mathbb{K} \text{swap}_c(A|e) \tag{4.115}$$

If: Suppose

$\mathbb{K} =$  (4.116)

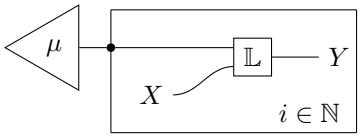
where μ is column exchangeable, and consider any two $x, x' \in X^{\mathbb{N}}$ such that some subsequences are equal $x_S = x'_T$ with $S, T \subset \mathbb{N}$ and $|S| = |T| = [n]$.

For any $\{A_i \in \mathcal{Y} | i \in S\}$, let $A_S = \text{swap}_{[n] \leftrightarrow S} \times_{i \in [n]} A_i \times Y^{\mathbb{N}}$, $A_T = \text{swap}_{S \leftrightarrow T}(A_S)$, $B = (x_i i)_{i \in S}$ and $C = (x_i i)_{i \in T} = (x_{\text{swap}_{S \leftrightarrow T}}(i) i)_{i \in S}$. By Equations 4.101 and 4.105

$$\mathbb{K}(A_S|x) = \mu \text{swap}_{S \leftrightarrow B}(A_S) \tag{4.117}$$
$$= \mu \text{swap}_{T \leftrightarrow C}(A_T) \qquad \text{by column exchangeability of } \mu \tag{4.118}$$
$$= \mathbb{K}(A_T|\text{swap}_{S \leftrightarrow T}(x)) \tag{4.119}$$
$$= \text{swap}_{S \leftrightarrow T} \mathbb{K}(A_T|x) \tag{4.120}$$
$$= \text{swap}_{S \leftrightarrow T} \mathbb{K} \text{swap}_{S \leftrightarrow T}(A_S|x) \tag{4.121}$$

so $\mathbb{K}$ is causally contractible by Theorem 4.3.5. □

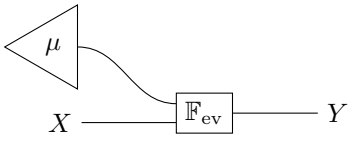
Theorem 4.3.7. *A kernel $\mathbb{K} : X^{\mathbb{N}} \nrightarrow Y^{\mathbb{N}}$ is causally contractible if and only if there exists some set H, $\mu \in \Delta(H)$ and $\mathbb{L} : H \times X \nrightarrow Y$ such that*

$\mathbb{K} =$  (4.122)

$$\iff \tag{4.123}$$

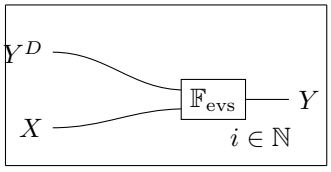
$$\mathbb{K}(\bigtimes_{i \in \mathbb{N}} A_i | (x_i)_{i \in \mathbb{N}}) = \int_H \prod_{i \in \mathbb{N}} \mathbb{L}(A_i|h, x_i) \mu(\mathrm{d}h) \tag{4.124}$$

Proof. By Theorem 4.3.6, we can represent the conditional probability $\mathbb{K}$ as

$\mathbb{K} =$  (4.125)

where μ is column exchangeable.

As a preliminary, we will show

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

(4.126)

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

$$(x, (y_i)_{i \in X}) \mapsto y_x \tag{4.127}$$

Recall that ev is the function

$$((x_i)_{\mathbb{N}}, (y_{ji})_{j,i \in X \times \mathbb{N}}) \mapsto (y_{x_i i})_{i \in \mathbb{N}} \tag{4.128}$$

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

$$\mathbb{F}_{\text{ev}}(\bigtimes_{i \in \mathbb{N}} A_i | (x_i)_{\mathbb{N}}, (y_{ji})_{i \in X \times \mathbb{N}}) = \delta_{(y_{x_i i})_{i \in \mathbb{N}}}(\bigtimes_{i \in \mathbb{N}} A_i) \tag{4.129}$$

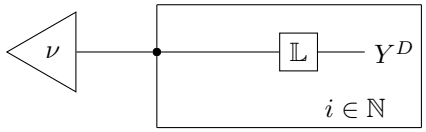
$$= \prod_{i \in \mathbb{N}} \delta_{y_{x_i i}}(A_i) \tag{4.130}$$

$$= \prod_{i \in \mathbb{N}} \mathbb{F}_{evs}(A_i | x_i, (y_{ji})_{j \in X}) \tag{4.131}$$

$$= \left(\bigotimes_{i \in \mathbb{N}} \mathbb{F}_{\text{evs}} \right) (\bigtimes_{i \in \mathbb{N}} A_i | (x_i)_{\mathbb{N}}, (y_{ji})_{j \in X \times \mathbb{N}}) \tag{4.132}$$

Which is what we wanted to show.

Only if: By the column exchangeability of μ, from Kallenberg (2005) we have some H, a directing random measure $\nu \in \Delta(H)$ and a kernel $\mathbb{L} : H \nrightarrow Y^X$ such that

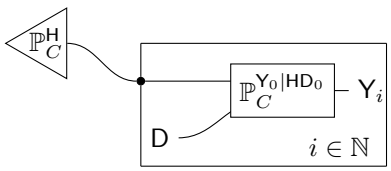
$\mu =$ 

(4.133)

$$\iff \tag{4.134}$$

$$\mu(\bigtimes_{i \in \mathbb{N}} A_i) = \int_H \prod_{i \in \mathbb{N}} \mathbb{L}(A_i | h) \mu(\mathrm{d}h) \qquad \forall A_i \in \mathcal{Y}^X \tag{4.135}$$

By Equations 4.125 and 4.126

$\mathbb{K} =$ 

(4.136)

Where we can connect the outputs of μ to the inputs of $\mathbb{F}_{\text{evs}}$ "inside the plate" as the plates in Equations 4.126 and 4.133 are equal in number and each connected wire represents a single copy of Y^D.

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

$$\mathbb{K}(\bigtimes_{i\in\mathbb{N}} A_i | x) = \int_H \prod_{i\in\mathbb{N}} \mathbb{L}(A_i|h, x_i)\mu(\mathrm{d}h) \tag{4.137}$$

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

$$\begin{aligned}
\mathbb{K}(A_S|x) &= \int_H \prod_{i\in S} \mathbb{L}(A_i|h, x_i)\mu(\mathrm{d}h) && (4.138)\\
&= \int_H \prod_{i\in T} \mathbb{L}(A_i|h, x_{\text{swap}_{S\leftrightarrow T}(i)})\mu(\mathrm{d}h) && (4.139)\\
&= \text{swap}_{S\leftrightarrow T}\mathbb{K}(A_T|x) && (4.140)\\
&= \text{swap}_{S\leftrightarrow T}\mathbb{K}\text{swap}_{S\leftrightarrow T}(A_S|x) && (4.141)
\end{aligned}$$

So by Theorem 4.3.5, $\mathbb{K}$ is causally contractible. □

Lemma 4.3.8 (Infinitely extended kernels)**.** *Given a collection of Markov kernels $\mathbb{K}_i : X^i \nrightarrow Y^i$ for all $i \in \mathbb{N}$, if we have for every $j > i$*

$$\mathbb{K}_j(id_{X_i} \otimes del_{X_{j-i}}) = \mathbb{K}_i \otimes del_{X_{j-i}} \tag{4.142}$$

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

$$\mathbb{K}(id_{X_i} \otimes del_{X_{j-i}}) = \mathbb{K}_i \otimes del_{X_{j-i}} \tag{4.143}$$

Proof. Take any $x \in X^{\mathbb{N}}$ and let $x_{|m} \in X^n$ be the first n elements of x. By Equation 4.142, for any $A_i \in \mathcal{Y}$, $i \in [m]$

$$\mathbb{K}_n(\bigtimes_{i\in[m]} A_i \times Y^{n-m} | x_{|n}) = \mathbb{K}_m(\bigtimes_{i\in[m]} A_i | x_{|m}) \tag{4.144}$$

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

$$\mathbb{K}_n\text{swap}_\rho(\bigtimes_{i\in[m]} A_{\rho(i)} \times Y^{n-m} | x_{|n}) = \mathbb{K}_n(\bigtimes_{i\in[m]} A_i \times Y^{n-m} | x_{|n}) \tag{4.145}$$

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

$$\mathbb{Q}_d(\bigtimes_{i\in[n]} A_i \times Y^{\mathbb{N}}) = \mathbb{K}_n(\bigtimes_{i\in[n]} A_{\rho(i)} | d_{|n}) \tag{4.146}$$

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

$$\mathbb{Q}_d(\bigtimes_{i\in[n]} A_i \times Y^{\mathbb{N}}) \geq \mathbb{Q}_d(\bigtimes_{i\in[p]} A_i \times Y^{\mathbb{N}}) \tag{4.147}$$

$$\geq \mathbb{Q}_d(\bigtimes_{i\in\mathbb{N}} A_i) \tag{4.148}$$

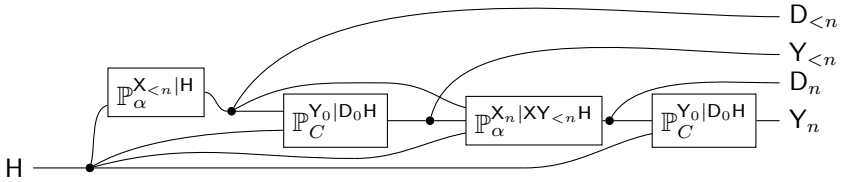
so by the Monotone convergence theorem, the sequence $\mathbb{Q}_d(\bigtimes_{i\in[n]} A_i)$ converges as $n \to \infty$ to $\mathbb{Q}_d(\bigtimes_{i\in\mathbb{N}} A_i)$. $d \mapsto \mathbb{Q}_d^{\mathsf{Z}_n}(\bigtimes_{i\in[n]} A_i)$ is measurable for all n, $\{A_i \in \mathcal{Y} | i \in \mathbb{N}\}$ by Equation 4.146, and so $d \mapsto Q_d$ is also measurable.

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

Theorem 4.3.9 (Conditional independence in augmented causally contractible model)**.** *Suppose we have a probability set $\mathbb{P}_C$ on Ω with causally contractible $\mathbb{N}$-comb $\mathbb{P}_C^{\mathsf{Y} \wr \mathsf{D}}$, $\mathsf{Y} = (\mathsf{Y}_i)_{i\in\mathbb{N}}$, $\mathsf{D} = (\mathsf{D}_i)_{i\in\mathbb{N}}$, augmented with the hypothesis variable H as in Lemma 4.3.11. Then for all n, $\mathsf{Y}_n \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{<n}, \mathsf{D}_{<n}) | (\mathsf{H}, \mathsf{D}_n)$.*

Proof. For all $\alpha \in C$, $n \in \mathbb{N}$

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



(4.149)

Thus for all α, by Theorem 2.4.27

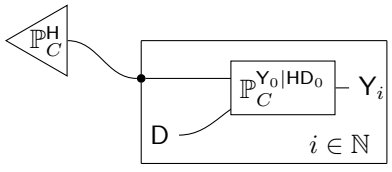
$$\mathbb{P}_C^{\mathsf{Y}_n|\mathsf{D}_n\mathsf{H}} \otimes \mathrm{del}_{D^{n-1}\times Y^{n-1}} \overset{\mathbb{P}_\alpha}{\cong} \mathbb{P}_C^{\mathsf{Y}_n|\mathsf{D}_{[n]}\mathsf{Y}_{<n}\mathsf{H}} \tag{4.150}$$

which implies $\mathsf{Y}_n \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{<n}, \mathsf{D}_{<n}) | (\mathsf{H}, \mathsf{D}_n)$. □

Theorem 4.3.10 (Conditional independence in non-data-dependant causally contractible model)**.** *Suppose we have a probability set $\mathbb{P}_C$ on Ω with causally contractible $\mathbb{N}$-comb $\mathbb{P}_C^{\mathsf{Y} \wr \mathsf{D}}$, $\mathsf{Y} = (\mathsf{Y}_i)_{i\in\mathbb{N}}$, $\mathsf{D} = (\mathsf{D}_i)_{i\in\mathbb{N}}$, augmented with the hypothesis variable H as in Lemma 4.3.11. Then for all n, $\mathsf{Y}_n \perp\!\!\!\perp^e_{\mathbb{P}_C} (\mathsf{Y}_{<n}, \mathsf{D}_{<n}) | (\mathsf{H}, \mathsf{D}_n)$.*

Lemma 4.3.11. *Suppose we have a probability set $\mathbb{P}'_C$ on $(\Omega', \mathcal{F}')$ with causally contractible $\mathbb{N}$-comb $\mathbb{P}_C^{\mathsf{Y}' \wr \mathsf{D}''}$, $\mathsf{Y}' = (\mathsf{Y}'_i)_{i\in\mathbb{N}}$, $\mathsf{D}' = (\mathsf{D}'_i)_{i\in\mathbb{N}}$. Then there exists H and an* augmented *model $\mathbb{P}_C$ on $(\Omega, \mathcal{F}) := ((\Omega' \times H, \mathcal{F}' \otimes \mathcal{H})$ such that $\mathbb{P}_C \Pi_{\Omega'} = \mathbb{P}'_C$ and, defining $\mathsf{H} : \Omega' \times H \to H$ as the projection onto H,*

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



(4.151)

Proof. Let H be the hypothesis space from Theorem 4.3.7 and $\mathbb{P}_C^{\mathsf{H}}$ be some directing random measure for $\mathbb{P}_C^{\prime\mathsf{Y}\prime\mathsf{D}}$. Then by Theorem 4.151, Equation 4.151 holds. Furthermore, by equality of combs, we have for all $\alpha \in C$

$$\mathbb{P}_\alpha^{\mathsf{YD}} = \mathbb{P}_\alpha^{\prime\mathsf{Y}'\mathsf{D}'} \tag{4.152}$$

Define $\mathsf{W}' : \Omega' \to \Omega'$ as the identity function on Ω', $\mathsf{W} : \Omega' \times H \to \Omega'$ as the projection to Ω'. For each $\alpha \in C$, define $\mathbb{P}_\alpha$ by

$$\mathbb{P}_\alpha^{\mathsf{W}} = \mathbb{P}_\alpha^{\mathsf{YD}} \odot \mathbb{P}_\alpha^{\prime\mathsf{W}'|\mathsf{Y}'\mathsf{D}'}(\mathrm{del}_{Y^{\mathbb{N}}D^{\mathbb{N}}H} \otimes \mathrm{id}_{\Omega'}) \tag{4.153}$$

Then

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathsf{W}} &= \mathbb{P}_\alpha \mathit{\Pi}_{\Omega'} && (4.154)\\
&= \mathbb{P}_\alpha^{\prime\mathsf{Y}'\mathsf{D}'} \odot \mathbb{P}_\alpha^{\prime\mathsf{W}'|\mathsf{Y}'\mathsf{D}'}(\mathrm{del}_{Y^{\mathbb{N}}D^{\mathbb{N}}H} \otimes \mathrm{id}_{\Omega'}) && (4.155)\\
&= \mathbb{P}_\alpha^{\prime\mathsf{W}'} && (4.156)
\end{aligned}$$

□

Chapter 5

Statistical Decision Theory

Can non-comb version of probability set represent statistical decision theory?

5.1 Summary

Statistical models are ubiquitous in the analysis of inference problems. A statistical model features a set of *states*, and each state is mapped to a probability distribution over *outcomes*. If we want to model problems involving *decisions* and *consequences*, we need to consider different kinds of statistical models. We introduce two types of model for this purpose: *two player statistical models* which differ from classical statistical model in that the state is assumed to consist of a decision and a *hypothesis* (the two players are the decision maker "player D" and the hypothesis selector "player H"). They model the consequences of decisions under various hypotheses. A *see-do model* is a special case of two player statistical model that can be used in situtations where some *observations* are available for review prior to selecting a decision. See-do models are the main focus of work here and problems involving observations, decisions and consequences will be discussed at length in Chapter 4.

A common simplifying assumption made when using classical statistical models is that they are *conditionally independent and identically distributed* (conditinally IID); this means that the model maps each state to an independent and identically distributed (IID) sequence of observations. This just a common choice, it is not a strict requirement. A similar assumption is likely to be useful for see-do models. If consequences depend on choices, then it does not make sense to assert that observations and consequences together form a single IID sequence of random variables, so we need to consider alternatives. We propose that models where observervations are an IID sequence and choices and consequences together are *independent and functionally identical* (IFI; defined later in this chapter) are similar to conditionally IID statistical models.

Instead of directly assuming that a conditionally IID model is appropriate, *De Finetti's representation theorem* shows that probability models where the sequence of observations is *exchangeable* induce conditionally IID statistical models. The assumption of exchangeabile observations is preferable to the assumption of IID observations if a probability model is being used to represent subjective uncertainty. We investigate whether there is an analogous result relating "exchangeability-like" assumptions for see-do models to "IID-like" assumptions. We show that there is: in particular, a see-do "forecast" with exchangeable observations and *functionally exchangeable* decision to consequence maps induces a see-do model with IID observations and IFI consequences.

The assumption of functional exchangeability will appear again in Chapter 6 as part of the definition of *counterfactual models*, and the joint assumptions of exchangeable observations and functionally exchangeable consequences to motivate the assumption of *imitability* in Chapter **??**, an assumption that in combination with a number of other assumptions allows for inference of consequences from data.

5.2 Modelling observations, choices and consequences

5.2.1 Modelling observations with statistical models

Statistical models are a ubiquitous type of model in statistics and machine learning. They consist of a set of *states* $(S, \mathcal{S})$, and for each state the model prescribes a single probability distribution on a given measurable set of *outcomes* $(O, \mathcal{O})$.

Definition 5.2.1 (Statistical model)**.** A statistical model is a set of states $(S, \mathcal{S})$, a set of outcomes $(O, \mathcal{O})$ and a stochastic map $\mathbb{T} : S \to \Delta(\mathcal{O})$.

Definition 5.2.2 (State and outcome variables)**.** Given a statistical model $(\mathbb{T}, (O, \mathcal{O}), (S, \mathcal{S}))$, define the *state variable* $\mathsf{S} : S \times O \to S$ as the projection from $S \times O \to S$ and define the *outcome variable* $\mathsf{O} : S \times O \to O$ as the projection onto O.

The common example of a potentially biased coin is modelled with a statistical model. We suppose our coin has some rate of heads $\theta \in [0, 1]$, and we furthermore suppose that for each θ the result of flipping the coin can be modeled (in some sense) by the probability distribution Bernoulli(θ). The statistical model here is the set of states $S = [0, 1]$ (corresponding to *rates of heads*), the observation space $O = \{0, 1\}^n$ with the discrete sigma-algebra (where n is the number of flips observed) and the stochastic map $\mathbb{B} : [0, 1] \to \Delta(\mathcal{P}(0, 1))$ which is given by $\mathbb{B} : \theta \to \text{Bernoulli}(\theta)$.

Almost any theoretical treatment of statistics or machine learning will at some point make use of statistical models to describe the problem they are addressing – for a collection of examples from the last 100 years, see Goodfellow

et al. (2016); Vapnik (2013); Bishop (2006); Le Cam (1996); Freedman (1963); Wald (1950); de Finetti ([1937] 1992); Fisher ([1925] 1992). They are often simply assumed without a great deal of discussion of why this type of model is chosen, or what role they play.

If we want to reason about how well some learning algorithm performs in some context, we typically require a reasonable model of the context in which the learning algorithm operates. The algorithms themselves may not give us such a model. Because learning almost always operates in a context with noise an uncertainty, we need models that can handle noise and uncertainty. Probability models are a very common choice for this. In addition, it is often assumed that we do not know with certainty the exact probability model that should be used to model a context. A statistical model assumes a certain number of states may prevail – reflecting uncertainty in the "mechanics of the world" – and given any state it gives us a probability distribution – reflecting uncertainty remaining after the mechanics of the world are well-understood.

Learning algorithms don't necessarily implement reasonable models of the world. For example, consider a linear regressor that takes a set of predictors $x \in X$ and targets $y \in Y$ and returns some $\beta \in B$ such that $(y - x^T\beta)^2$ is as small as possible. It is possible to interpret B as a set of states, and consider the learner to be implementing the statistical model $(\mathbb{T}, B, \mathcal{L}_{X\to Y})$ where $\mathcal{L}_{X\to Y}$ is the set of liner function $X \to Y$ given by $\{x \mapsto x^T\beta | \beta \in B\}$ and $\mathbb{T}$ maps a state $\beta \in B$ deterministically to the function $x \mapsto x^T\beta$. This is, formally, a statistical model, but it is not one that would typically be considered a good model of the world in the kinds of problems that a linear regressor is used to solve. One problem with this model is that it is deterministic - the outcome for any $\beta \in B$ will be a particular function $X \to Y$. However, it will almost never be the case that some set of targets y will be an exact function of some set of predictors x, and insisting on an exact functional relationship will typically give very poor generalisation results if this demand can be satisfied at all.

Suppose we want to ask whether the function f given by a linear regressor is useful for some purpose. In order to address this question, we want to consider a more appropriate model of the world than the crude statistical model given above, and consider what behaviour we will see from the regressor under different assumptions imposed on this model. Statistical models typically are used to serve the purpose of a "more appropriate model of the world". In this example, we might consider a statistical model $(\mathbb{T}, H, O)$ where for each $h \in H$, $\mathbb{T}_h \in \Delta(\mathcal{Y}\otimes\mathcal{X})$ such that $\mathbb{T}_h^{\mathsf{Y}|\mathsf{X}} = \text{Normal}(\mathsf{X}^T\beta_h, \sigma_h)$. If we assume the data generating process is described by such a probability distribution for some h, we can ask questions like "does the linear regressor output a β such that $\mathsf{Y} - \mathsf{X}^T\beta < \epsilon$ with high probability for all $h \in H$?"

5.2.2 Modelling choices and consequences with two-player statistical models

The states in a statistical model are usually considered to be "under the control of nature". In the possibly biased coin example above, if we were to consider some "player D" acutally flipping the coin and trying to infer the bias, we would typically assume that their opinion about the coin's bias does not affect the coin's actual bias; they could decide it is biased towards heads when in fact it is completely unbiased. In some cases player D can make choices that affect the outcomes. Suppose player D has the option to choose how high to toss the coin – perhaps they can aim for a toss height anywhere from 10 to 50cm. This plausibly affects the outcomes of their coin toss and, unlike the coin's bias, they gets to choose the height they intend to toss it. If they decide to toss it to a height of 15cm then 15cm is the height they have chosen to toss it to. Unlike the state, which can differ from whatever player D ultimately decides on, the choice made by player D is the same thing as whatever they ultimately decide on. We call features of the state that are not under player D's control *hypotheses* and features that are under player D's control *decisions*, and statistical models in which the state is the Cartesian product of a set of hypotheses and a set of decisions "two player statistical models" (the two players being nature or "player H" and the decision maker or "player D").

Definition 5.2.3 (Two player statistical model)**.** A *two-player statistical model* is a tuple $(\mathbb{T}, \mathsf{H}, \mathsf{D}, \mathsf{O})$ where $(\mathbb{T}, (H \times D, \mathcal{H} \otimes \mathcal{D}), (O, \mathcal{O}))$ is a statistical model and $\mathsf{H} : H \times D \times O \to H$, $\mathsf{D} : H \times D \times O \to D$ and $\mathsf{O} : H \times D \times O \to O$ are measurable functions that project elements of $H \times D \times O$ to their respective codomains. H is called the *hypothesis*, D the *decision* and O the *outcome*.

Whenever we propose a two player statistical model, we will also assume for any random variables $\mathsf{X} : H \times D \times O \to X$ and $\mathsf{Y} : H \times D \times O \to Y$, a disintegration $\mathbb{K}^{\mathsf{Y}|\mathsf{XDH}} : X \times D \times H \to \Delta(\mathcal{Y})$ exists (see Theorem **??** for a sufficient condition).

The problems that we will mostly study in this work, in addition to having a second player ("player D"), will often involve some data X that is observed before the second player is able to make a choice. Two player statistical models with *observations* are called *see-do models*.

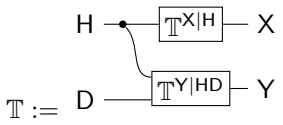
Definition 5.2.4 (See-Do model)**.** A *see-do model* $(\mathbb{T}, \mathsf{H}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ is a two-player statistical model along with two additional random variables: the *observation* $\mathsf{X} : H \times D \times O \to X$ and the *consequence* $\mathsf{Y} : H \times D \times O \to Y$. The outcome variable is defined to be the coupled product of the observation and the consequence $\mathsf{O} = (\mathsf{X}, \mathsf{Y})$, and we will leave this implicit when specifying a see-do model. A see-do model must observe the conditional independence:

$$\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D} | \mathsf{H} \tag{5.1}$$

We can informally read the independence requirement as saying "the observations are independent of the decision given the hypothesis". This does not

imply that probability models we construct from $\mathbb{T}$ will necessarily have the property that D and X will be independent conditional on H, and in fact this will often not be the case. In Chapter 6 we will argue that this requirement captures the intuition that observations are not "affected" by decisions. For now, we will observe that this independence requirement means that $\mathbb{T}$ can be drawn with no path from D to X.

Explicitly, the independence on line 5.1 implies that the kernel $\mathbb{T}$ can be drawn as follows:



(5.2)

In this picture, again informally, Y takes input from D but X does not.

Chapter 6

See-do models, interventions and counterfactuals

6.1 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

6.2 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

6.3 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

6.4 Unit Potential Outcomes models

- Counterfactual random variables Yx 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 Ruben, 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 develope here is surprisingly appliccable 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 theorm is appliccable 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 Interention 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 $\mathbb{T} : D \times \mathsf{H} \to \Delta(\mathcal{E})$ 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 cruicially 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).

6.4.1 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 $D_0 = \{\text{flip the switch}, \text{don't flip the switch}\}$. 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:

$D_1 :=\{$"walk to the switch and press it with my thumb$''$,
"trip over the lego on the floor, hop to the light switch and stab my finger at it$''$,
"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 :={"walk to the switch and press it with my thumb″,
"trip over the lego on the floor, hop to the light switch and stab my finger at it″,
"stay in bed″,
"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 $\mathbb{T} : H \times D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$, define the *consequence model* $\mathbb{C} : H \times D \to \Delta(\mathcal{Y})$ as $\mathbb{C} := \mathbb{T}^{\mathsf{Y}|\mathsf{HD}}$.

Definition 6.4.1 (D-causation)**.** Given a hypothesis $h \in H$ and a consequence model $\mathbb{C} : H \times D \to \Delta(\mathcal{Y})$, random variables $\mathsf{Y}_1 : Y \times D \to Y_1$, $\mathsf{Y}_2 : Y \times D \to Y_2$ and $\mathsf{D} : Y \times D \to D$ (defined the usual way), Y_1 D-causes Y_2 iff $\mathsf{Y}_2 \perp\!\!\!\perp_{\mathbb{C}} \mathsf{D}|\mathsf{Y}_1\mathsf{H}$.

6.4.2 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 6.4.2 (Limited unresponsiveness)**.** Given states S, deterministic consequence models $\mathbb{C}_s : D \to \Delta(F)$ for each $s \in A$ and a random variables $\mathsf{Y}_1 : F \to Y_1$, $\mathsf{Y}_2 : F \to Y_2$, Y_1 is unresponsive to D in states limited by Y_2 if $\mathbb{C}^{\mathsf{Y}_2|\mathsf{SD}}_{(s,d)} = \mathbb{C}^{\mathsf{Y}_2|\mathsf{SD}}_{(s,d')} \implies \mathbb{C}^{\mathsf{Y}_1|\mathsf{SD}}_{(s,d)} = \mathbb{C}^{\mathsf{Y}_1|\mathsf{SD}}_{(s,d')}$ for all $d, d' \in D$, $s \in S$. Write $\mathsf{Y}_1 \not\leftharpoondown_{\mathsf{Y}_2} \mathsf{D}$

Lemma 6.4.3 (Limited unresponsiveness implies D-causation)**.** *For deterministic consequence models,* $\mathsf{Y}_1 \not\leftharpoondown_{\mathsf{Y}_2} \mathsf{D}$ *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 $\mathbb{C}^{\mathsf{Y}_1\mathsf{Y}_2|\mathsf{SD}}_{s,d} = \delta_{y_1(s,d)} \otimes \delta_{y_2(s,d)}$.

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 $f : Y_2 \times S \to Y_1$ 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 $\mathbb{M} : Y_2 \times S \times D \to \Delta(\mathcal{Y}_1)$ by $(y_2, s, d) \mapsto \delta_{f(y_2,s)}$. $\mathbb{M}$ is a version of $\mathbb{C}^{\mathsf{Y}_1|\mathsf{Y}_2,\mathsf{S},\mathsf{D}}$ because, for all $A \in \mathcal{Y}_2$, $B \in \mathcal{Y}_1$, $s \in S$, $d \in D$:

$$\mathbb{C}^{\mathsf{Y}_2|\mathsf{SD}}_{(s,d)}\curlyvee(\mathbb{M}\otimes \mathrm{Id}) = \int_A \mathbb{M}(y_2', d, s; B)d\delta_{y_2(s,d)}(y_2') \tag{6.1}$$

$$= \int_A \delta_{f(y_2',s)}(B)d\delta_{y_2(s,d)}(y_2') \tag{6.2}$$

$$= \delta_{f(y_2(s,d),s)}(B)\delta_{y_2(s,d)}(A) \tag{6.3}$$

$$= \delta_{y_1(s,d)}(B)\delta_{y_2(s,d)}(A) \tag{6.4}$$

$$= \delta_{y_2(s,d)} \otimes \delta_{y_1(s,d)}(A \times B) \tag{6.5}$$

$\mathbb{M}$ is clearly constant in D. Therefore $\mathsf{Y}_1 \perp\!\!\!\perp_{\mathbb{C}} \mathsf{D}|\mathsf{Y}_2\mathsf{S}$. □

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, $\mathsf{F} \in \{0, 1\}$ is the final position of the switch and $\mathsf{L} \in \{0, 1\}$ is the final state of the light. We have

define this

$$\mathbb{C}^{\mathsf{LF}|\mathsf{DS}}_{d,[i,m]} = \delta_{(d \text{ XOR } i) \text{ AND } m} \otimes \delta_{(d \text{ XOR } i) \text{ AND } m} \tag{6.6}$$

Within states $[0, 0]$ and $[1, 0]$, the light is always off, so $\mathsf{F} = a \implies \mathsf{L} = 0$ for any a. In states $[0, 1]$ and $[1, 1]$, $\mathsf{F} = 1 \implies \mathsf{L} = 1$ and $\mathsf{F} = 0 \implies \mathsf{L} = 0$. Thus $\mathsf{L} \not\rightsquigarrow_{\mathsf{F}} \mathsf{D}$. However, suppose we take a mixture of consequence models:

$$\mathbb{C}_\gamma = \frac{1}{4}\mathbb{C}_{\cdot,[0,0]} + \frac{1}{4}\mathbb{C}_{\cdot,[0,1]} + \frac{1}{2}\mathbb{C}_{\cdot,[1,1]} \tag{6.7}$$

$$\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma = \frac{1}{4}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \otimes \begin{bmatrix}1 & 0\\ 1 & 0\end{bmatrix} + \frac{1}{4}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \otimes \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} \otimes \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} \tag{6.8}$$

Then

$$[1, 0]\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma = \frac{1}{4}[0, 1] \otimes [1, 0] + \frac{1}{4}[0, 1] \otimes [0, 1] + \frac{1}{2}[1, 0] \otimes [1, 0] \tag{6.9}$$

$$[1, 0]\curlyvee(\mathbb{C}^{\mathsf{F}|\mathsf{D}}_\gamma \otimes \mathbb{C}^{\mathsf{L}|\mathsf{D}}_\gamma) = (\frac{1}{2}[0, 1] + \frac{1}{2}[1, 0]) \otimes (\frac{1}{4}[0, 1] + \frac{3}{4}[1, 0]) \tag{6.10}$$

$$\implies [1, 0]\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma \neq [1, 0]\curlyvee(\mathbb{C}^{\mathsf{F}|\mathsf{D}}_\gamma \otimes \mathbb{C}^{\mathsf{L}|\mathsf{D}}_\gamma) \tag{6.11}$$

Thus under the prior γ, 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 $\mathsf{H} \in S$, 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 $\mathsf{H} \in S$, and in this way we differ substantially from Heckerman and Shachter (1995).

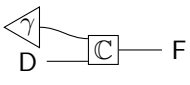
define this

Instead, we simply extend the definition of D-causation to mixtures of hypotheses: if $\gamma \in \Delta(\mathsf{H})$ is a mixture of hypotheses, define $\mathbb{C}_\gamma := (\gamma \otimes \mathbf{Id})\mathbb{C}$. Then X D-causes Y relative to γ iff $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{C}_\gamma} \mathsf{D}|\mathsf{X}$.

Theorem 6.4.4 shows that under some conditions, D-causation can hold for arbitrary mixtures over subsets of the hypothesis class H.

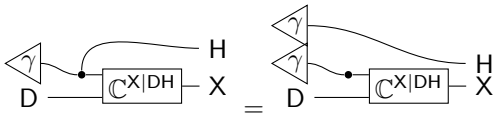
Theorem 6.4.4 (Universal D-causation)**.** *If* $\mathsf{X} \perp\!\!\!\perp \mathsf{H}|\mathsf{D}$ *for all* $\mathsf{H}, \mathsf{H}' \in S \subset \mathsf{H}$ *and* X *D-causes* Y *in all* $\mathsf{H} \in S$*, then* X *D-causes* Y *with respect to all mixed consequence models* $\mathbb{C}_\gamma$ *for all* $\gamma \in \Delta(\mathsf{H})$ *with* $\gamma(S) = 1$*.*

Proof. For $\gamma \in \Delta(\mathsf{H})$, define the mixture

$\mathbb{C}_\gamma :=$ 

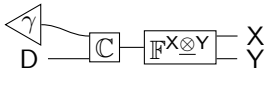
(6.12)

Because $\mathbb{C}_{\mathsf{H}}^{\mathsf{X}|\mathsf{D}} = \mathbb{C}_{\mathsf{H}'}^{\mathsf{X}|\mathsf{D}}$ for all $\mathsf{H}, \mathsf{H}' \in \mathsf{H}$, we have

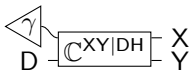


(6.13)

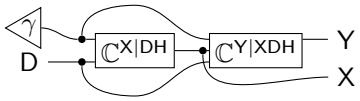
Also

$\mathbb{C}_\gamma^{\mathsf{XY}|\mathsf{D}} =$ 

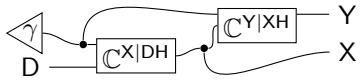
(6.14)

$=$ 

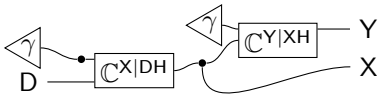
(6.15)

$=$ 

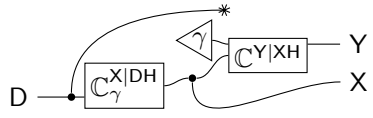
(6.16)

$\overset{\mathsf{Y} \perp\!\!\!\perp \mathsf{D}|\mathsf{XH}}{=}$ 

(6.17)

$\overset{6.13}{=}$ 

(6.18)

$\overset{6.13}{=}$ 

(6.19)

Equation 6.19 establishes that $(\gamma \otimes \mathbf{Id}_X \otimes {}^{*}\!\uparrow_D)\mathbb{C}^{\mathsf{Y}|\mathsf{XH}}$ is a version of $\mathbb{C}_\gamma^{\mathsf{Y}|\mathsf{XD}}$, and thus $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{C}_\gamma} \mathsf{D}|\mathsf{X}$.

This can also be derived from the semi-graphoid rules:

$$\begin{aligned}
\mathsf{H} \perp\!\!\!\perp \mathsf{D} \wedge \mathsf{H} \perp\!\!\!\perp \mathsf{X}|\mathsf{D} &\implies \mathsf{H} \perp\!\!\!\perp \mathsf{XD} && (6.20)\\
&\implies \mathsf{H} \perp\!\!\!\perp \mathsf{D}|\mathsf{X} && (6.21)\\
\mathsf{D} \perp\!\!\!\perp \mathsf{H}|\mathsf{X} \wedge \mathsf{D} \perp\!\!\!\perp \mathsf{Y}|\mathsf{XH} &\implies \mathsf{D} \perp\!\!\!\perp \mathsf{Y}|\mathsf{X} && (6.22)\\
&\implies \mathsf{Y} \perp\!\!\!\perp \mathsf{D}|\mathsf{X} && (6.23)
\end{aligned}$$

□

6.4.3 Properties of D-causation

If X D-causes Y relative to $\mathbb{C}_\mathsf{H}$, then the following holds:

$$\mathbb{C}_\mathsf{H}^{\mathsf{X}|\mathsf{D}} = \quad \mathsf{D} \;\text{—}\boxed{\mathbb{C}^{\mathsf{X}|\mathsf{D}}}\text{—}\boxed{\mathbb{C}^{\mathsf{Y}|\mathsf{X}}}\text{—}\; \mathsf{Y} \tag{6.24}$$

This follows from version (2) of Definition **??**:

$$\mathbb{C}_\mathsf{H}^{\mathsf{X}|\mathsf{D}} = \quad \mathsf{D} \;\text{—}\bullet\text{—}\boxed{\mathbb{C}^{\mathsf{X}|\mathsf{D}}}\text{—}\boxed{\mathbb{C}^{\mathsf{Y}|\mathsf{XD}}}\text{—}\; \mathsf{Y} \tag{6.25}$$

$$= \quad \mathsf{D} \;\text{—}\bullet\text{—}\boxed{\mathbb{C}^{\mathsf{X}|\mathsf{D}}}\text{—}\boxed{\mathbb{C}^{\mathsf{Y}|\mathsf{X}}}\text{—}\; \mathsf{Y} \tag{6.26}$$

$$= \quad \mathsf{D} \;\text{—}\boxed{\mathbb{C}^{\mathsf{X}|\mathsf{D}}}\text{—}\boxed{\mathbb{C}^{\mathsf{Y}|\mathsf{X}}}\text{—}\; \mathsf{Y} \tag{6.27}$$

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?

6.4.4 Decision sequences and parallel decisions

Just as observations X can be a sequence of random variables X_1, X_2, ..., D can be a sequence of "sub-choices" D_1, D_2, ... Note that by positing such a sequence there is no requirement that D_1 comes "before" D_2 in any particular sense.

6.5 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 $\{V_1, V_2, ..., V_n\}$ 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 $\{f_1, f_2, ..., f_n\}$ such that each f_i is a mapping from (the respective domains of) $U_i \cup PA_i$ to V_i, where $Ui \subseteq U$ 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

$$v_i = f_i(pa_i, u_i), \qquad i \in 1, ...n,$$

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 $h : \mathcal{M} \times D \to \mathcal{M}$, 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 $Y.(\cdot) : D \times U \to Y$ (distinct from Y, which appears to be a function $U \to$ something) and calls $Y.(\cdot)$ the "potential response". We could always consider the variable $\mathsf{V} := \underline{\otimes}_{i\in[n]}\mathsf{V}_i$ and define the "total potential response" $\mathbf{g} := \mathsf{V}.(\cdot)$, 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 $\mathbf{g} : D \times U \to V$.

Pearl's model also features a second deterministic function $\mathbf{f} : U \to Y$, and G is derived from F via the equation modifications permitted by D. It is straightforward to show that an arbitrary function $\mathbf{f} : U \to Y$ 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 $D \times U \to Y$ (though many possible choices would be quite unusual).

Pearl's counterfactual model seems to essentially be a deterministic map $\mathbf{g} : D \times U \to V$ 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 $D \to \Delta(\mathcal{V})$, 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 represents 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.

References

A. V. Banerjee, S. Chassang, and E. Snowberg. Chapter 4 - Decision Theoretic Approaches to Experiment Design and External ValidityaaWe thank Esther Duflo for her leadership on the handbookăand 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.

Christopher Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag, New York, 2006. ISBN 978-0-387-31073-2. URL https://www.springer.com/gp/book/9780387310732.

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.

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.

R.P. Feynman. *The Feynman lectures on physics*. Le cours de physique de Feynman. Intereditions, France, 1979. ISBN 978-2-7296-0030-3. INIS Reference Number: 13648743.

R. A. Fisher. Statistical Methods for Research Workers. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, Springer Series in Statistics, pages 66–70. Springer, New York, NY, [1925] 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_6. URL https://doi.org/10.1007/978-1-4612-4380-9_6.

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.

David A. Freedman. On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case. *Annals of Mathematical Statistics*, 34(4):1386–1403, December 1963. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177703871. URL https://projecteuclid.org/euclid.aoms/1177703871. Publisher: Institute of Mathematical Statistics.

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.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

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.

Miguel A. Hernán and James M Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020. URL https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

Eric Horvitz, David Heckerman, and Curtis Langlotz. A Framework for Comparing Alternative Formalisms for Plausible Reasoning. January 1986. URL https://openreview.net/forum?id=rJNeX0gdbr.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, 2015. ISBN 978-0-521-88588-1. doi: 10.1017/CBO9781139025751. URL https://www.cambridge.org/core/books/causal-inference-for-statistics-social-and-biomedical-sciences/71126BE90C58F1A431FE9B2DD07938AB.

Bart Jacobs, Aleks Kissinger, and Fabio Zanasi. Causal Inference by String Diagram Surgery. In Mikoaj Bojaczyk and Alex Simpson, editors, *Foundations of Software Science and Computation Structures*, Lecture Notes in Computer Science, pages 313–329. Springer International Publishing, 2019. ISBN 978-3-030-17127-8.

Richard C. Jeffrey. *The Logic of Decision*. University of Chicago Press, July 1965. ISBN 978-0-226-39582-1.

Olav Kallenberg. The Basic Symmetries. In *Probabilistic Symmetries and Invariance Principles*, Probability and Its Applications, pages 24–68. Springer, New York, NY, 2005. ISBN 978-0-387-28861-1. doi: 10.1007/0-387-28861-9_2. URL https://doi.org/10.1007/0-387-28861-9_2.

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.

L. Le Cam. Comparison of Experiments - A Short Review.pdf. *IMS Lecture Notes - Monograph Series*, 30, 1996.

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.

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 cigarettelung cancer link: evidentiary traditions, corporate denial, global toll. *Tobacco Control*, 21(2):87–91, March 2012. ISSN 0964-4563, 1468-3318. doi: 10.1136/tobaccocontrol-2011-050338. URL https://tobaccocontrol.bmj.com/content/21/2/87. Publisher: BMJ Publishing Group Ltd Section: The shameful past.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov Properties for Acyclic Directed Mixed Graphs. *arXiv:1701.06686 [stat]*, January 2017. URL http://arxiv.org/abs/1701.06686. arXiv: 1701.06686.

Donald B. Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, March 2005. ISSN 0162-1459. doi: 10.1198/016214504000001880. URL https://doi.org/10.1198/016214504000001880.

L. J. Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46:55–67, 1951. ISSN 1537-274X(Electronic),0162-1459(Print). doi: 10.2307/2280094.

Leonard J. Savage. *The Foundations of Statistics*. Wiley Publications in Statistics, 1954.

P. Selinger. A Survey of Graphical Languages for Monoidal Categories. In Bob Coecke, editor, *New Structures for Physics*, Lecture Notes in Physics, pages 289–355. Springer, Berlin, Heidelberg, 2011. ISBN 978-3-642-12821-9. doi: 10.1007/978-3-642-12821-9_4. URL https://doi.org/10.1007/978-3-642-12821-9_4.

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

Ilya Shpitser and Judea Pearl. Complete Identification Methods for the Causal Hierarchy. *Journal of Machine Learning Research*, 9(Sep):1941–1979, 2008. ISSN ISSN 1533-7928. URL https://www.jmlr.org/papers/v9/shpitser08a.html.

Brian Skyrms. Causal Decision Theory. *The Journal of Philosophy*, 79(11):695–711, November 1982. doi: 10.2307/2026547. URL https://www.pdcnet.org/pdc/bvdb.nsf/purchase?openform&fp=jphil&id=jphil_1982_0079_0011_0695_0711.

Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

Statista. Cigarettes - worldwide | Statista Market Forecast, 2020. URL https://www.statista.com/outlook/50010000/100/cigarettes/worldwide.

Katie Steele and H. Orri Stefánsson. Decision Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2020 edition, 2020. URL https://plato.stanford.edu/archives/win2020/entries/decision-theory/.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer Science & Business Media, June 2013. ISBN 978-1-4757-3264-1. Google-Books-ID: EqgACAAAQBAJ.

J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Theory of games and economic behavior. Princeton University Press, Princeton, NJ, US, 1944.

N. N. Vorobev. Consistent Families of Measures and Their Extensions. *Theory of Probability & Its Applications*, 7(2), 1962. doi: 10.1137/1107014. URL http://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=tvp&paperid=4710&option_lang=eng.

Abraham Wald. *Statistical decision functions*. Statistical decision functions. Wiley, Oxford, England, 1950.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.

Robert Wiblin. Why smoking in the developing world is an enormous problem and how you can help save lives, 2016. URL https://80000hours.org/problem-profiles/tobacco/.

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: