# Causal Statistical Decision Theory|What are interventions?

David Johnston

July 29, 2021

# 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:

> $\mathsf{X}_i$ is a *cause* of $\mathsf{X}_j$ if there is an *ideal intervention* on $\mathsf{X}_i$ that changes the value $\mathsf{X}_j$

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 $\mathsf{U}$ contains $\mathsf{X}_i$, $\mathsf{X}_j$ and "all intervenable variables that *cause*" both $\mathsf{X}_i$ and $\mathsf{X}_j$ [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

Todo: conditional expectation, martingale convergence

Almost sure equality convention

Existence of disintegrations for choosing probability measures

We use probability theory extensively to develop the theory of decision-theoretic causal inference. This chapter introduces key results and conventions which will be used in later chapters. Many of the results are well-known, and some are extensions of well-known results to Markov kernel spaces which are a generalisation of probability spaces (and a special case of conditional probability spaces).

On the other hand, we also make extensive use of a string diagram notation for Markov kernels, which is not well-known so even readers who are highly familiar with probability theory. The introduction of string diagrams begins in Section 2.0.3.

### 2.0.1 Probability Theory

Given a set $A$, a $\sigma$-algebra $\mathcal{A}$ is a collection of subsets of $A$ where

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

A measurable space $(A, \mathcal{A})$ is a set $A$ along with a $\sigma$-algebra $\mathcal{A}$. Sometimes the sigma algebra will be left implicit, in which case $A$ will just be introduced as a measurable space.

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

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

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

Let $T$ be all the open subsets of $\mathbb{R}$. Then $\mathcal{B}(\mathbb{R}) := \sigma(T)$ is the *Borel $\sigma$-algebra* on the reals. This definition extends to an arbitrary topological space $A$ with topology $T$.

A *standard measurable set* is a measurable set $A$ that is isomorphic either to a discrete measurable space $A$ or $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For any $A$ that is a complete separable metric space, $(A, \mathcal{B}(A))$ is standard measurable.

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 $\{A_i\} \subset \mathcal{E}$, $\mu(\cup_i A_i) = \sum_i \mu(A_i)$

Write by $\Delta(\mathcal{E})$ the set of all probability measures on $\mathcal{E}$.

A particular probability measure we will often discuss is the *Dirac measure*. For any $x \in X$, the Dirac measure $\delta_x \in \Delta(\mathcal{X})$ is the probability measure where $\delta_x(A) = 0$ if $x \notin A$ and $\delta_x(A) = 1$ if $x \in A$.

Given another measurable space $(F, \mathcal{F})$, a *stochastic map* or *Markov kernel* is a map $\mathbb{M} : E \times \mathcal{F} \to [0, 1]$ such that

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

Extending the subscript notation, for $\mathbb{C} : X \times Y \to \Delta(\mathcal{Z})$ and $x \in X$ we will write $\mathbb{C}_{x,\cdot}$ for the "curried" map $y \mapsto \mathbb{C}_{x,y}$. If $\mathbb{C}$ is a Markov kernel with respect to $(X \times Y, \mathcal{X} \otimes \mathcal{Y}), (Z, \mathcal{Z})$ then it is straightforward to show that $\mathbb{C}_{x,\cdot}$ is a Markov kernel with respect to $(Y, \mathcal{Y}), (Z, \mathcal{Z})$.

This yields the notational conventions for arbitrary kernel $\mathbb{C}$:

- $\mathbb{C}$ with no subscripts is a Markov kernel
- $\mathbb{C}_{\cdot,a,b}$ with at least one $\cdot$ subscript is a Markov kernel
- $\mathbb{C}_y$ with no $\cdot$ subscripts is a probability measure

The map $x \mapsto \mathbb{M}_x$ is of type $E \to \Delta(\mathcal{F})$. We will abuse notation somewhat to write $\mathbb{M} : E \to \Delta(\mathcal{F})$. In this sense, we view Markov kernels as maps from elements of $E$ to probability measures on $\mathcal{F}$. This is simply a convention that helps us to think about constructions involving Markov kernels, and it is equally valid to view Markov kernels as maps from elements of $\mathcal{F}$ to measurable functions $E \to [0, 1]$, a view found in Clerc et al. (2017), or simply in terms of their definition above.

Given an indiscrete measurable space $(\{*\}, \{\{*\}, \emptyset\})$, we identify Markov kernels $\mathbb{N} : \{*\} \to \Delta(\mathcal{E})$ with the probability measure $\mathbb{N}_*$. In addition, there is a unique Markov kernel $* : E \to \Delta(\{\{*\}, \emptyset\})$ given by $x \mapsto \delta_*$ for all $x \in E$ which we will call the "discard" map.

Two Markov kernels $\mathbb{M}X \to \Delta(\mathcal{Y})$ and $\mathbb{N} : X \to \Delta(\mathcal{Y})$ are equal iff for all $x \in X$, $A \in \mathcal{Y}$

$$\mathbb{M}_x(A) = \mathbb{N}_x(A) \tag{2.1}$$

We will typically be more concerned with "almost sure" equality than exact equality, which will be defined later.

### 2.0.2 Product Notation

Probability measures, Markov kernels and measurable functions can be combined to yield new probability measures, Markov kernels or measurable functions. Given $\mu \in \Delta(\mathcal{X})$, $\mathsf{T} : Y \to T$, $\mathbb{M} : X \to \Delta(\mathcal{Y})$ and $\mathbb{N} : Y \to \Delta(\mathcal{Z})$ define:

The **measure-kernel** product $\mu\mathbb{M} : \mathcal{Y} \to [0, 1]$ where for all $A \in \mathcal{Y}$,

$$\mu\mathbb{M}(A) := \int_X \mathbb{M}_x(A) d\mu(x) \tag{2.2}$$

The **kernel-function** product $\mathbb{M}\mathsf{T} : X \to T$ where for all $x \in X$:

$$\mathbb{M}\mathsf{T}(x) := \int_Y T(y) d\mathbb{M}_x(y) \tag{2.3}$$

The **kernel-kernel** product $\mathbb{M}\mathbb{N} : X \to \Delta(\mathcal{Z})$ where for all $x \in X$, $A \in \mathcal{Z}$:

$$(\mathbb{M}\mathbb{N})_x(A) := \int_Y \mathbb{N}_y(A) d\mathbb{M}_x(y) \tag{2.4}$$

All kernel products are associative (Çinlar, 2011). An intuition for this notation can be gained from thinking of probability measures $\mu \in \Delta(\mathcal{X})$ as row vectors, Markov kernels $\mathbb{M}, \mathbb{N}$ as matrices and measurable functions $\mathsf{T} : Y \to T$ as column vectors and kernel products are vector-matrix and matrix-matrix products. If the $X, Y, Z$ and $T$ are discrete spaces then this analogy is precise.

Finally, the **tensor product** $\mathbb{M} \otimes \mathbb{N} : X \times Y \to \Delta(\mathcal{Y} \otimes \mathcal{Z})$ is yields the kernel that applies $\mathbb{M}$ and $\mathbb{N}$ "in parallel". For all $x \in X$, $y \in Y$, $G \in \mathcal{Y}$ and $H \in \mathcal{Z}$:

$$(\mathbb{M} \otimes \mathbb{N})_{x,y}(G \times H) := \mathbb{M}_x(G)\mathbb{N}_y(H) \tag{2.5}$$

### 2.0.3 String Diagrams

Some constructions are unwieldly in product notation; for example, given $\mu \in \Delta(\mathcal{E})$ and $\mathbb{M} : E \to (\mathcal{F})$, it is not straightforward to write an expression using kernel products and tensor products that represents the "joint distribution" given by $A \times B \mapsto \int_A \mathbb{M}(x; B) d\mu$.

An alternative notation known as *string diagrams* provides greater expressive capability than product notation while being more visually clear than integral notation. Cho and Jacobs (2019) provides an extensive introduction to string diagram notation for probability theory.

Key features of string diagrams include:

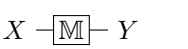
- String diagrams as they are used in this work can always be interpreted as a mixture of kernel-kernel products and tensor products of Markov kernels
- String diagrams are the subject of a coherence theorem: two string diagrams that differ only by planar deformation are always equal (Selinger, 2010). This also holds for a number of additional transformations detailed below
  - Informally, diagrams that look like they should be the same are in fact the same

#### Elements of string diagrams

The basic elements of a string diagram are Markov kernels. Diagrams representing Markov kernels can be assembled into larger diagrams by taking regular products or tensor products.

Indiscrete spaces play a key role in string diagrams. An indiscrete space is any one element measurable space $(\{*\}, \{\emptyset, \{*\}\})$ which admits the unique probability measure $\mu : \{\emptyset, \{*\}\} \to (0, 1)$ given by $\mu(\emptyset) = 0$, $\mu(\{*\}) = 1$. Any probability measure $\mu \in \Delta(\mathcal{X})$ can be interpreted as a Markov kernel $\mu' : \{*\} \to \Delta(\mathcal{X})$ where $\mu'_* = \mu$ (note that $*$ is the *only* argument $\mu'$ can be given).

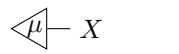
A Markov kernel $\mathbb{M} : X \to \Delta(\mathcal{Y})$ can always be represented as a rectangular box with input and output wires labeled with the relevant spaces:




(2.6)

Note that we will later substitute labelling wires with spaces for labelling them with random variable names.

Probability measures $\mu \in \Delta(\mathcal{X})$ can be written as triangles:



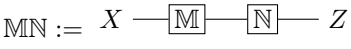

(2.7)

We can exploit the identification of the probability measure $\mu$ with the Markov kernel $\mu' : \{*\} \to \Delta(\mathcal{X})$ given by $* \mapsto \mu$ to preserve the principle that any element of a string diagram is a Markov kernel. Under this identification, all elements of string diagrams are Markov kernels. Because, furthermore, and the

set of Markov kernels is closed under the product and tensor product operations introduced below, a consequence is that all well-formed string diagrams are Markov kernels.

Cho and Jacobs (2019) defines the operation of *conditioning* using kernel-function products which makes use of kernel-function products which and, unlike measure-kernel products, kernel-function products do not in general produce Markov kernels. At this stage, we do not make use of a graphical conditioning operation, but we note that this could be an useful direction to extend the graphical theory presented here.

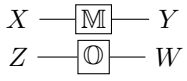
**Elementary operations** Kernel-kernel products have a visually similar representations in string diagram notation to the previously introduced product notation. Given $\mathbb{M} : X \to \Delta(\mathcal{Y})$ and $\mathbb{N} : Y \to \Delta(\mathcal{Z})$, we have

$\mathbb{M}\mathbb{N} :=$  (2.8)

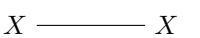

For $\mu \in \Delta(\mathcal{E})$,

$\mu\mathbb{M} :=$  (2.9)

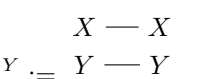

Tensor products in string diagram notation are represented by vertical juxtaposition. For $\mathbb{O} : Z \to \Delta(\mathcal{W})$:

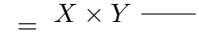
$\mathbb{M} \otimes \mathbb{O} :=$  (2.10)


A space $X$ is identified with the identity kernel $\mathrm{Id}^X : X \to \Delta(\mathcal{X})$, $x \mapsto \delta_x$. A bare wire represents an identity kernel or, equivalently, the space given by its labels:

$\mathrm{Id}^X :=$  (2.11)


Product spaces $X \times Y$ are identified with tensor products of identity kernels $X \times Y \cong \mathbb{I}^X \otimes \mathbb{I}^Y$. These can be represented either by two parallel wires or by a single wire equipped with appropriate labels:

$X \times Y \cong \mathrm{Id}^X \otimes \mathrm{Id}^Y :=$  (2.12)


$=$  (2.13)


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:


(2.14)

$\equiv$ (2.15)


(2.16)

**Markov kernels with special notation** A number of Markov kernels are given special notation distinct from the generic "box" above. This notation facilitates intuitive visual representation.

As has already been noted, the identity kernel $\mathbf{Id} : X \to \Delta(X)$ maps a point $x$ to the measure $\delta_x$ that places all mass on the same point:

$$\mathbf{Id} : x \mapsto \delta_x \equiv \quad X \text{ — } X \tag{2.17}$$

The identity kernel is an identity under left and right products:

$$\begin{aligned}
(\mathbb{K}\mathbf{Id})_w(A) &= \int_X \mathbf{Id}_x(A) d\mathbb{K}_w(x) && (2.18)\\
&= \int_X \delta_x(A) d\mathbb{K}_w(x) && (2.19)\\
&= \int_A d\mathbb{K}_w(x) && (2.20)\\
&= \mathbb{K}_w(A) && (2.21)\\
(\mathbf{Id}\mathbb{K})_w(A) &= \int_X \mathbb{K}_x(A) d\mathbf{Id}_w(x) && (2.22)\\
&= \int_X \mathbb{K}_x(A) d\delta_w(x) && (2.23)\\
&= \mathbb{K}_w(A) && (2.24)
\end{aligned}$$

The copy map $\curlyvee : X \to \Delta(\mathcal{X} \times \mathcal{X})$ maps a point $x$ to two identical copies of x:

$$\curlyvee : x \mapsto \delta_{(x,x)} \equiv \quad X \text{ —}\langle \begin{matrix} X \\ X \end{matrix} \tag{2.25}$$

The copy map "copies" its arguments to kernels or under the right product:

$$\begin{aligned}
\int_( X \times X) \mathbb{K}_{x',x''}(A) d\curlyvee_x(x', x'') &= \int_( X \times X) \mathbb{K}_{x',x''}(A) d\delta_{(x,x)}(x', x'') && (2.26)\\
&= \mathbb{K}_{x,x}(A) && (2.27)
\end{aligned}$$

The swap map $\rho : X \times Y \to \Delta(\mathcal{Y} \otimes \mathcal{X})$ swaps its inputs:

$$\rho := (x, y) \to \delta_{(y,x)} \equiv \begin{matrix} Y \\ X \end{matrix} \asymp \begin{matrix} X \\ Y \end{matrix} \tag{2.28}$$

Under products are taken with the swap map, arguments are interchanged. For $\mathbb{K} : X \times Y \to \Delta(\mathcal{Z})$ and $\mathbb{L} : Z \to \Delta(\mathcal{X} \otimes \mathcal{Y})$, $A \in \mathcal{X}$, $B \in \mathcal{Y}$:

$$(\rho\mathbb{K})_{y,x}(A) = \int_{(} X \times Y)\mathbb{K}_{x',y'}(A)d\rho_{(y,x)}(x', y') \quad = \int_{(} X \times Y)\mathbb{K}_{x',y'}(A)d\delta_{(x,y)}(x', y') \tag{2.29}$$

$$= \mathbb{K}_{x,y}(A) \tag{2.30}$$

$$(\mathbb{L}\rho)_z(B \times A) = \int_{X \times Y} \rho_{x',y'}(B \times A)d\mathbb{L}_z(x', y') \tag{2.31}$$

$$= \int_{X \times Y} \delta_{(y',x')}(B \times A)d\mathbb{L}_z(x', y') \tag{2.32}$$

$$= \mathbb{L}_z(A \times B) \tag{2.33}$$

The discard map $* : X \to \Delta(\{*\})$ maps every input to $\delta_*$, the unique probability measure on the indiscrete set $\{\emptyset, \{*\}\}$.

$$* : x \mapsto \delta_* \equiv \begin{matrix} X \\ \end{matrix} \longrightarrow * \tag{2.34}$$

Any measurable function $g : W \to X$ has an associated Markov kernel $\mathbb{F}^g : W \to \Delta(\mathcal{X})$ given by $\mathbb{F}^g : w \mapsto \delta_{g(w)}$. Given a probability measue $\mu \in \Delta(\mathcal{W})$, $\mu g$ is a measure-function product while $\mu\mathbb{F}^g$ is commonly called the pushforward measure $g_\#\mu$. We will generalise this slightly to the notion of *pushforward kernels*.

**Definition 2.0.1** (Kernel associated with a function)**.** Given a measurable function $g : W \to X$, define the function induced kernel $\mathbb{F}^g : W \to \Delta(\mathcal{X})$ to be the the Markov kernel $w \mapsto \delta_{g(w)}$ for all $w \in W$.

**Definition 2.0.2** (Pushforward kernel)**.** Given a kernel $\mathbb{M} : V \to \Delta(\mathcal{W})$ and a measurable function $g : W \to X$, the *pushforward kernel* $g_\#\mathbb{M} : V \to \Delta(\mathcal{X})$ is the kernel $g_\#\mathbb{M}$ such that $(g_\#\mathbb{M})_a(B) = \mathbb{M}_a(g^{-1}(B))$ for all $a \in V$, $B \in \mathcal{X}$.

**Lemma 2.0.3** (Pushforward kernels are functional kernel products)**.** *Given a kernel $\mathbb{M} : V \to \Delta(\mathcal{W})$ and a measurable function $g : W \to X$, $g_\#\mathbb{M} = \mathbb{M}\mathbb{F}^g$.*

*Proof.* for any $a \in V$, $B \in \mathcal{X}$:

$$(\mathbb{M}\mathbb{F}^g)_a(B) = \int_W \delta_{g(y)}(B) d\mathbb{M}_a(y) \tag{2.35}$$

$$= \int_W \delta_y(g^{-1}(B)) d\mathbb{M}_a(y) \tag{2.36}$$

$$= \int_{g^{-1}(B)} d\mathbb{M}_a(y) \tag{2.37}$$

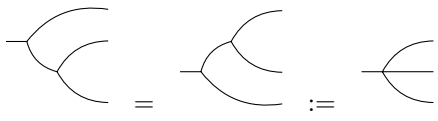
$$= (g_\# \mathbb{M})_a(B) \tag{2.38}$$

□

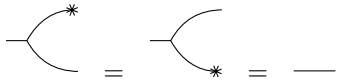
**Working With String Diagrams**

There are a relatively small number of manipulation rules that are useful for string diagrams. In addition, we will define graphically analogues of the standard notions of *conditional probability*, *conditioning*, and infinite sequences of exchangeable random variables.

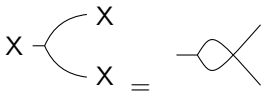
**Axioms of Symmetric Monoidal Categories** For the following, we either omit labels or label diagrams with their domain and codomain spaces, as we are discussing identities of kernels rather than identities of components of a condtional probability space. Recalling the unique Markov kernels defined above, the following equivalences, known as the *commutative comonoid axioms*, hold among string diagrams:



(2.39)

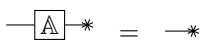


(2.40)



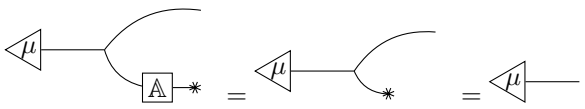

(2.41)

The discard map $*$ can "fall through" any Markov kernel:




(2.42)

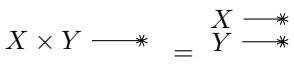
Combining 2.40 and 2.42 we can derive the following: integrating $\mathbb{A} : X \to \Delta(\mathcal{Y})$ with respect to $\mu \in \Delta(\mathcal{X})$ and then discarding the output of $\mathbb{A}$ leaves us with $\mu$:



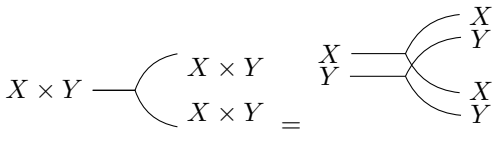

(2.43)

In elementary notation, this is equivalent to the fact that, for all $B \in \mathcal{X}$, $\int_B \mathbb{A}(x; B) d\mu(x) = \mu(B)$.

The following additional properties hold for $*$ and $\curlyvee$:

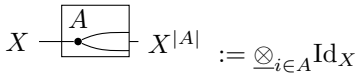



(2.44)



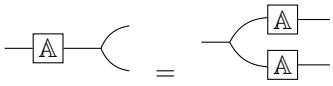

(2.45)

Note that for some set $X$, the copy map $\curlyvee_X = \mathrm{Id}_X \underline{\otimes} \mathrm{Id}_X$. This combined with 2.41 allows us to define the $A$-copy map for some $A \subseteq \mathbb{N}$:




$:= \underline{\otimes}_{i \in A} \mathrm{Id}_X$ (2.46)

A key fact that *does not* hold in general is




(2.47)

In fact, it holds only when $\mathbb{A}$ is a *deterministic* kernel.

**Definition 2.0.4** (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 $x \in E$, $B \in \mathcal{F}$.

**Theorem 2.0.5** (Copy map commutes for deterministic kernels (Fong, 2013))**.** *Equation 2.47 holds iff* $\mathbb{A}$ *is deterministic.*

**Examples**

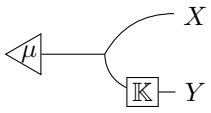
Given $\mu \in \Delta(X), \mathbb{K} : X \to \Delta(Y)$, $A \in \mathcal{X}$ and $B \in \mathcal{Y}$:

$$A \times B \mapsto \int_A \mathbb{K}(x; B) d\mu(x) \tag{2.48}$$

$$\equiv \tag{2.49}$$

$$\mu \gtrdot (\mathbf{Id}_X \otimes \mathbb{K}) \tag{2.50}$$

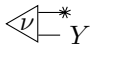
$$\equiv \tag{2.51}$$




(2.52)

Cho and Jacobs (2019) calls this operation "integrating $\mathbb{K}$ with respect to $\mu$".

Given $\nu \in \Delta(\mathcal{X} \otimes \mathcal{Y})$, define the marginal $\nu^{\mathsf{Y}} \in \Delta(\mathcal{Y}) : B \mapsto \mu(X \times B)$ for $B \in \mathcal{Y}$. Say that $\nu^{\mathsf{Y}}$ is obtained by marginalising over "$X$" (a notion that can be made more precise by assigning names to wires). Then




$$\nu(* \otimes \mathrm{Id}^Y) = \tag{2.53}$$

$$\nu(* \otimes \mathrm{Id}^Y)(B) := \nu(* \otimes \mathrm{Id}^Y)(B \times \{*\}) \tag{2.54}$$

$$= \int_{X \times Y} \mathrm{Id}^Y_y(B) *_x(\{*\}) d\nu(x, y) \tag{2.55}$$

$$= \int_{X \times Y} \delta_y(B) \delta_*(\{*\}) d\nu(x, y) \tag{2.56}$$

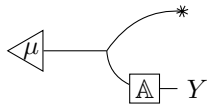
$$= \int_{X \times B} d\nu(x, y) \tag{2.57}$$

$$= \nu(X \times B) \tag{2.58}$$

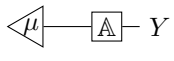
$$= \nu^{\mathsf{Y}}(B) \tag{2.59}$$

Thus the action of the erasing wire "$X$" is equivalent to marginalising over "$X$".

Consider the result of marginalising 2.52 over "$X$":




$$\nu^Y(B) = \tag{2.60}$$




$$= \tag{2.61}$$

### 2.0.4 Random Variables

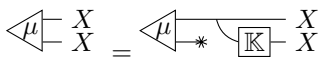
The summary of this section is:

- Random variables are usually defined as measurable functions on a *probability space*
- It's possible to define them as measurable functions on a *Markov kernel space* instead
- It is useful to label wires with random variable names instead of names of spaces

Probability theory is primarily concerned with the behaviour of *random variables*. This behaviour can be analysed via a collection of probability measures and Markov kernels representing joint, marginal and conditional distributions of random variables of interest. In the framework developed by Kolmogorov, this collection of joint, marginal and conditional distributions is modeled by a single underlying *probability space*, and random variables by measurable functions on the probability space.

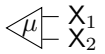
We use the same approach here, with a couple of additions. We are interested in variables whose outcomes depend both on random processes and decisions. Suppose that given a particular distribution over decision variables, a probability distribution over the decision variables and random variables is obtained. Such a model is described by a Markov kernel rather than a probability distribution. We therefore investigate *Markov kernel spaces*.

In the graphical notation that we are using, random variables can be thought of as a means of assigning unambiguous names to each wire in a set of diagrams. In order to do this, it is necessary to suppose that all diagrams in the set describe properties of an *ambient Markov kernel* or *ambient probability measure*. Consider the following example with the ambient probability measure $\mu \in \Delta(\mathcal{X} \otimes \mathcal{X})$. Suppose we have a Markov kernel $\mathbb{K} : X \to \Delta(\mathcal{X})$ such that the following holds:




(2.62)

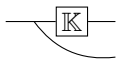
Suppose that we also assign the names $\mathsf{X}_1$ to the upper output wire and $\mathsf{X}_2$ to the lower output wire in the diagram of $\mu$:




(2.63)

Then it seems sensible to call $\mathbb{K}$ "the probability of $\mathsf{X}_2$ given $\mathsf{X}_1$". We will make this precise, and it will match the usual notion of the probability of one variable given another (see Çinlar (2011) for a definition of this usual notion).

**Definition 2.0.6** (Probability space, Markov kernel space)**.** A *Markov kernel space* $(\mathbb{K}, (D, \mathcal{D}), (\Omega, \mathcal{F}))$ is a Markov kernel $\mathbb{K} : D \to \Delta(\mathcal{D} \otimes \mathcal{F})$, called the *ambient kernel*, along with the sample space $(\Omega, \mathcal{F})$ and the domain $(D, \mathcal{D})$. Define the *canonical extension* $\mathbb{K}^*$ of $\mathbb{K}$ such that

$\mathbb{K}^* :=$ 

(2.64)

For brevity, we will omit the $\sigma$-algebras in further definitions of Markov kernel spaces: $(\mathbb{K}, D, \Omega)$.

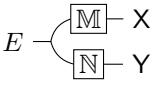
A *probability space* $(\mathbb{P}, \Omega, \mathcal{F})$ is a probability measure $\mathbb{P} : \Delta(\Omega)$, which we call the *ambient measure*, along with the *sample space* $\Omega$ and the *events* $\mathcal{F}$. A probability space is equivalent to a Markov kernel space with domain $D = \{*\}$ - note that $\Omega \times \{*\} \cong \Omega$.

**Definition 2.0.7** (Random variable)**.** Given a Markov kernel space $(\mathbb{K}, D, \Omega)$, a random variable $\mathsf{X}$ is a measurable function $\Omega \times D \to E$ for arbitrary measurable $E$.

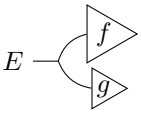
**Definition 2.0.8** (Domain variable)**.** Given a Markov kernel space $(\mathbb{K}, D, \Omega)$, the *domain variable* $\mathsf{D} : \Omega \times D \to D$ is the distinguished random variable $\mathsf{D} : (x, d) \mapsto d$.

Unlike random variables on probability spaces, random variables on Markov kernel spaces do not generally have unique marginal distributions. An analogous operation of *marginalisation* can be defined, and the result is in general a Markov kernel. We will define marginalisation via coupled tensor products.

**Definition 2.0.9** (Coupled tensor product $\underline{\otimes}$)**.** Given two Markov kernels $\mathbb{M}$ and $\mathbb{N}$ or functions $f$ and $g$ with shared domain $E$, let $\mathbb{M} \underline{\otimes} \mathbb{N} := \curlyvee(\mathbb{M} \otimes \mathbb{N})$ and $f \underline{\otimes} g := \curlyvee(f \otimes g)$ where these expressions are interpreted using standard product notation. Graphically:

$\mathbb{M} \underline{\otimes} \mathbb{N} :=$ 

(2.65)

$f \underline{\otimes} g :=$ 

(2.66)

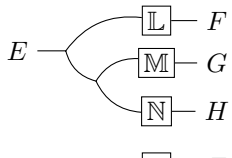
The operation denoted by $\underline{\otimes}$ is associative (Lemma 2.0.10), so we can without ambiguity write $f \underline{\otimes} g \underline{\otimes} h = (f \underline{\otimes} g) \underline{\otimes} h = f \underline{\otimes} (g \underline{\otimes} h)$ for finite groups of functions or Markov kernels sharing a domain.

The notation $\underline{\otimes}_{i \in [N]} f_i$ means $f_1 \underline{\otimes} f_2 \underline{\otimes} ... \underline{\otimes} f_N$. This is unambiguous due to Lemma 2.0.10

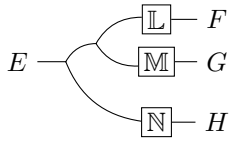
**Lemma 2.0.10** ($\underline{\otimes}$ is associative)**.** *For Markov kernels* $\mathbb{L} : E \to \delta(\mathcal{F})$*,* $\mathbb{M} : E \to \delta(\mathcal{G})$ *and* $\mathbb{N} : E \to \delta(\mathcal{H})$*,* $(\mathbb{L} \underline{\otimes} \mathbb{M}) \underline{\otimes} \mathbb{N} = \mathbb{L} \underline{\otimes} (\mathbb{M} \underline{\otimes} \mathbb{N})$.

*Proof.*

$$\mathbb{L} \underline{\otimes} (\mathbb{M} \underline{\otimes} \mathbb{N}) =$$




(2.67)




$$= \tag{2.68}$$

$$= (\mathbb{L} \underline{\otimes} \mathbb{M}) \underline{\otimes} \mathbb{N} \tag{2.69}$$

This follows directly from Equation 2.39. □

**Definition 2.0.11** (Marginal distribution, marginal kernel)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ and the random variable $\mathsf{X} : \Omega \to G$ the *marginal distribution* of $\mathsf{X}$ is the probability measure $\mathbb{P}^{\mathsf{X}} := \mathbb{P}\mathbb{F}^{\mathsf{X}}$.

See Lemma 2.0.3 for the proof that this matches the usual definition of marginal distribution.

Given a Markov kernel space $(\mathbb{K}, \Omega, \mathcal{F}, D, \mathcal{D})$ and the random variable $\mathsf{X} : \Omega \to G$, the *marginal kernel* is $\mathbb{K}^{\mathsf{X}|\mathsf{D}} := \mathbb{K}^* \mathbb{F}^{\mathsf{X}}$. Recall that $\mathbb{K}^*$ is the canonical extension of $\mathbb{K}$ (Definition 2.0.6)

**Definition 2.0.12** (Joint distribution, joint kernel)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ and the random variables $\mathsf{X} : \Omega \to G$ and $\mathsf{Y} : \Omega \to H$, the *joint distribution* of $\mathsf{X}$ and $\mathsf{Y}$, $\mathbb{P}^{\mathsf{XY}} \in \Delta(\mathcal{G} \otimes \mathcal{H})$, is the marginal distribution of $\mathsf{X} \underline{\otimes} \mathsf{Y}$. That is, $\mathbb{P}^{\mathsf{XY}} := \mathbb{P}\mathbb{F}^{\mathsf{X} \underline{\otimes} \mathsf{Y}}$

This is identical to the definition in Çinlar (2011) if we note that the random variable $(\mathsf{X}, \mathsf{Y}) : \omega \mapsto (\mathsf{X}(\omega), \mathsf{Y}(\omega))$ (Çinlar's definition) is precisely the same thing as $\mathsf{X} \underline{\otimes} \mathsf{Y}$.

Analogously, the joint kernel $\mathbb{K}^{\mathsf{XY}|\mathsf{D}}$ is the product $\mathbb{K}^* \mathbb{F}^{\mathsf{X} \underline{\otimes} \mathsf{Y}}$.

Joint distributions and kernels have a nice visual representation, as a result of Lemma 2.0.13 which follows.

**Lemma 2.0.13** (Product marginalisation interchange)**.** *Given two functions, the kernel associated with their coupled product is equal to the coupled product of the kernels associated with each function.*

*Given* $\mathsf{X} : \Omega \to G$ *and* $\mathsf{Y} : \Omega \to H$, $\mathbb{F}^{\mathsf{X} \underline{\otimes} \mathsf{Y}} = \mathbb{F}^{\mathsf{X}} \underline{\otimes} \mathbb{F}^{\mathsf{Y}}$

*Proof.* For $a \in \Omega$, $B \in \mathcal{G}$, $C \in \mathcal{H}$,

$$\mathbb{F}^{\mathsf{X} \underline{\otimes} \mathsf{Y}}(a; B \times C) = \delta_{\mathsf{X}(a), \mathsf{Y}(a)}(B \times C) \tag{2.70}$$

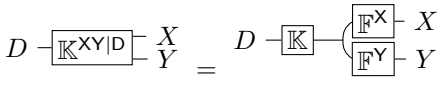
$$= \delta_{\mathsf{X}(a)}(B) \delta_{\mathsf{Y}(a)}(C) \tag{2.71}$$

$$= (\delta_{\mathsf{X}(a)} \otimes \delta_{\mathsf{Y}(a)})(B \times C) \tag{2.72}$$

$$= \mathbb{F}^{\mathsf{X}} \underline{\otimes} \mathbb{F}^{\mathsf{Y}} \tag{2.73}$$

Equality follows from the monotone class theorem. □

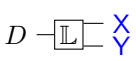
**Corollary 2.0.14.** *Given a Markov kernel space* $(\mathbb{K}, \Omega, D)$ *and random variables* $\mathsf{X} : \Omega \times D \to X$, $\mathsf{Y} : \Omega \times D \to Y$, *the following holds:*




(2.74)

We will now define wire labels for "output" wires.

**Definition 2.0.15** (Wire labels - joint kernels)**.** Suppose we have a Markov kernel space $(\mathbb{K}, D, \Omega)$, random variables $\mathsf{X} : \Omega \times D \to X$, $\mathsf{Y} : \Omega \times D \to Y$ and a Markov kernel $\mathbb{L} : D \to \Delta(\mathcal{X} \times \mathcal{Y})$. The following *output labelling* of $\mathbf{L}$:

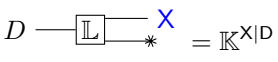



(2.75)

is *valid* iff

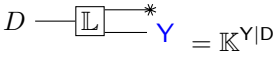
$$\mathbb{L} = \mathbb{K}_{\mathsf{XY|D}} \tag{2.76}$$

and



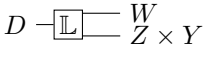

(2.77)

and




(2.78)

The second and third conditions are nontrivial: suppose $\mathsf{X}$ takes values in some product space $\mathit{Range}(\mathsf{X}) = W \times Z$, and $\mathsf{Y}$ takes values in $Y$. Then we could have $\mathbb{L} = \mathbb{K}^{\mathsf{XY|D}}$ and draw the diagram

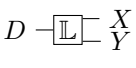



(2.79)

For *this* diagram, properties 2.77 and 2.78 do not hold, even though 2.76 does.

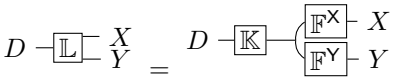
**Lemma 2.0.16** (Output label assignments exist)**.** *Given Markov kernel space* $(\mathbb{K}, D, \Omega)$*, random variables* $\mathsf{X} : \Omega \times D \to X$ *and* $\mathsf{Y} : \Omega \times D \to Y$ *then there exists a diagram of* $\mathbb{L} := \mathbb{K}^{\mathsf{XY|D}}$ *with a valid output labelling assigning* $\mathsf{X}$ *and* $\mathsf{Y}$ *to the output wires.*

*Proof.* By definition, $\mathbb{L}$ has signature $D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$. Thus, by the rule that tensor product spaces can be represented by parallel wires, we can draw



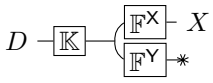

(2.80)

By Corollary 2.0.14, we have



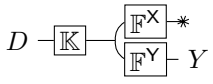

(2.81)

Therefore




$$= \mathbb{K}\mathbb{F}^{\mathsf{X}} \tag{2.82}$$

$$= \mathbb{K}^{\mathsf{X}|\mathsf{D}} \tag{2.83}$$



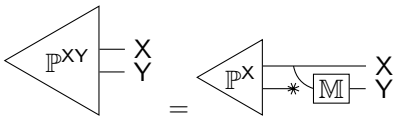

$$= \mathbb{K}\mathbb{F}^{\mathsf{Y}} \tag{2.84}$$

$$= \mathbb{K}^{\mathsf{Y}|\mathsf{D}} \tag{2.85}$$

□

In all further work, wire labels will be used without special colouring.

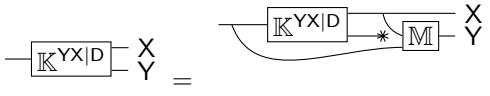
**Definition 2.0.17** (Disintegration)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$, and random variables $\mathsf{X}$ and $\mathsf{Y}$, we say that $\mathbb{M} : E \to \Delta(\mathcal{F})$ is a $\mathsf{Y}$ *given* $\mathsf{X}$ *disintegration* of $\mathbb{P}$ iff




(2.86)

$\mathbb{M}$ is a version of $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$, "the probability of $\mathsf{Y}$ given $\mathsf{X}$". Let $\mathbb{P}^{\{\mathsf{Y}|\mathsf{X}\}}$ be the set of all kernels that satisfy 2.86 and $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ an arbitrary member of $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$.

Given a Markov kernel space $(\mathbb{K}, D, \Omega)$ and random variables $\mathsf{X} : \Omega \times D \to X$, $\mathsf{Y} : \Omega \times D \to Y$, $\mathbb{M} : D \times E \to \Delta(\mathcal{F})$ is a $\mathsf{Y}$ *given* $\mathsf{DX}$ *disintegration* of $\mathbb{K}^{\mathsf{YX}|\mathsf{D}}$ iff



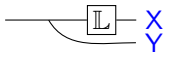

(2.87)

Write $\mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$ for the set of kernels satisfying 2.87 and $\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}$ for an arbitrary member of $\mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$.

**Definition 2.0.18** (Wire labels – input)**.** An input wire is *connected* to an output wire if it is possible to trace a path from the start of the input wire to the end of the output wire without passing through any boxes, erase maps or right facing triangles.

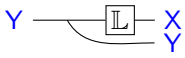
If an input wire is connected to an output wire and that output wire has a valid label $\mathsf{X}$, then it is valid to label the input wire with $\mathsf{X}$.

For example, if the following are valid output labels with respect to $(\mathbb{P}, \Omega)$:




(2.88)

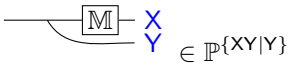
i.e. if $\mathbb{L} \in \mathbb{P}^{\{\mathsf{XY}|\mathsf{Y}\}}$, then the following is a valid input label:




(2.89)

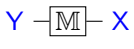
An input wire in a diagram for $\mathbb{M}$ may be labeled $\mathsf{X}$ *if and only if* copy and identity maps can be inserted to yield a diagram in which the input wire labeled $\mathsf{X}$ is connected to an output wire with valid label $\mathsf{X}$.

So, if $\mathbb{M} \in \mathbb{P}^{\{\mathsf{X}|\mathsf{Y}\}}$, then it is straightforward to show that




(2.90)

and hence the output labels are valid. Diagram 2.90 is constructed by taking the product of the copy map with $\mathbb{M} \otimes \mathbf{Id}$. Thus it is valid to label $\mathbb{M}$ with

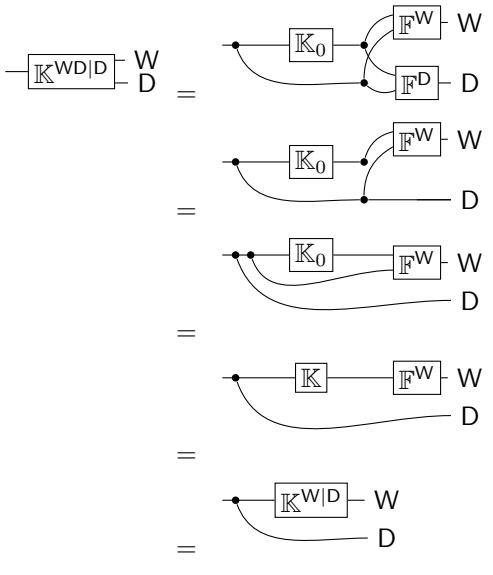



(2.91)

**Lemma 2.0.19** (Labeling of disintegrations)**.** *Given a kernel space* $(\mathbb{K}, D, \Omega)$*, random variables* $\mathsf{X}$ *and* $\mathsf{Y}$*, domain variable* $\mathsf{D}$ *and disintegration* $\mathbb{L} \in \mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$*, there is a diagram of* $\mathbb{L}$ *with valid input labels* $\mathsf{X}$ *and* $\mathsf{D}$ *and valid output label* $\mathsf{Y}$*.*

*Proof.* Note that for any variable $\mathsf{W} : \Omega \times D \to W$ and the domain variable

$\mathsf{D} : \Omega \times D \to D$ we have by definition of $\mathbb{K}$:

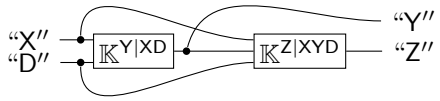



(2.92) (2.93) (2.94) (2.95) (2.96)

□

We use the informal convention of labelling wires in quote marks "$\mathsf{X}''$ if that wire is "supposed to" carry the label $\mathsf{X}$ but the label may not be valid.

**Lemma 2.0.20** (Iterated disintegration)**.** *Given a kernel space* $(\mathbb{K}, D, \Omega)$*, random variables* $\mathsf{X}$*,* $\mathsf{Y}$ *and* $\mathsf{Z}$ *and domain variable* $\mathsf{D}$*,*




$$\overset{a.s.}{=} \mathbb{K}^{\mathsf{YZ|XD}} \tag{2.97}$$

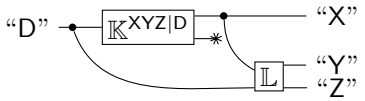
*Equivalently, for* $d \in D$ *and* $x \in X$*,* $A \in \mathcal{Y}$*,* $B \in \mathcal{Z}$*,*

$$(d, x; A \times B) \mapsto \int_A \mathbb{K}^{\mathsf{Z|XYD}}_{(x,y,d)}(B) d\mathbb{K}^{\mathsf{Y|XD}}_{(x,d)}(dy) \overset{a.s.}{=} \mathbb{K}^{\mathsf{YZ|XD}} \tag{2.98}$$

*Proof.* Define

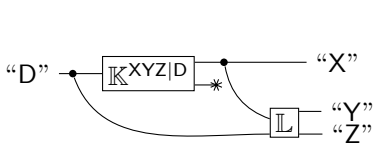
$$\mathbb{L} := (d, x; A \times B) \mapsto \int_A \mathbb{K}^{\mathsf{Z|XYD}}_{(x,y,d)}(B) d\mathbb{K}^{\mathsf{Y|XD}}_{(x,d)}(dy) \tag{2.99}$$

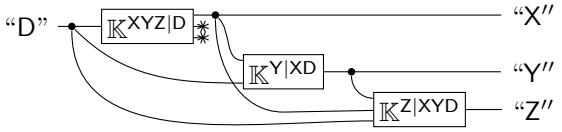
We need to show that $\mathbb{L}$ satisfies the disintegration property. That is

$\mathbb{K}^{\mathsf{XYZ|D}} =$ 

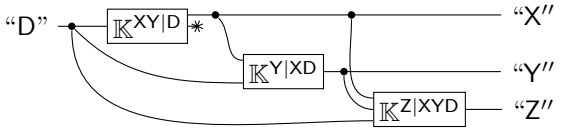

(2.100)

Substituting in the definition of $\mathbb{L}$:

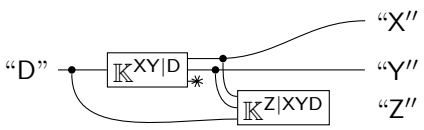



$=$ 


(2.101)

$=$ 


(2.102)

$=$ 


(2.103)

$= \mathbb{K}^{\mathsf{XYZ|D}}$ (2.104)

As required.

□

**Lemma 2.0.21.** *Suppose a kernel space $(\mathbb{K}, D, \Omega)$ for which disintegrations exist and a set of random variables $\{\mathsf{X}_i : D \times \Omega \to X_i | i \in [n]\}$, $\mathsf{X}_{[n]} := \underline{\otimes}_{i \in A} \mathsf{X}_i$ and $\mathsf{Y} : D \times \Omega \to \mathsf{Y}$ such that $\mathbb{K}^{\mathsf{X}_A|\mathsf{Y}}$ exists. Consider an arbitrary non-repeating sequence $[\mathsf{X}_{i_1}, \mathsf{X}_{i_2}, ..., \mathsf{X}_{i_n}]$ where $i_j = \rho(j)$ for some permuation $\rho : [n] \to [n]$ and a corresponding sequence*

$$[\mathsf{W}_{i_0}, \mathsf{W}_{i_1}, \mathsf{W}_{i_2}, ...\mathsf{W}_{i_{n-1}}] := [*, \underline{\otimes}_{j \in \{i_1\}} \mathsf{X}_j, \underline{\otimes}_{j \in \{i_1, i_2\}} \mathsf{X}_j, ..., \underline{\otimes}_{j \in \{i_1, i_2, ..., i_{n-1}\}} \mathsf{X}_j]$$

*Then $\mathbb{K}^{\mathsf{X}_A|\mathsf{Y}}$ can be almost surely uniquely defined by the collection of kernels $\mathbb{K}^{\mathsf{X}_{i_j}|\mathsf{W}_{i_{j-1}}\mathsf{Y}}$, $i \in A$, and in particular for any $\times_{j \in [n]} B_{i_j}$, $B_{i_j} \in \mathcal{X}_{i_j}$:*

$$\mathbb{K}_y^{\mathsf{X}_{\rho([n])}|\mathsf{Y}}(\times_{j \in [n]} B_{i_j}) \overset{a.s.}{=} \int_{B_{i_1}} ... \int_{B_{i_{n-1}}} \mathbb{K}^{\mathsf{X}_{i_n}|\mathsf{W}_{i_{n-1}}\mathsf{Y}}_{x_{i_1}, ..., x_{i_{n-1}}, y}(B_{i_n}) ... \mathbb{K}_y^{\mathsf{X}_1|\mathsf{Y}}(dx_{i_1})$$

(2.105)

*Proof.* By application of Lemma 2.0.20

$$\mathbb{K}_y^{\mathsf{X}_{\rho([n])}|\mathsf{Y}}(\times_{j\in[n]}B_{i_j}) = \int_{B_{i_1}} \mathbb{K}_{x_{i_1},y}^{\mathsf{X}_{\rho(\{2,..,n\})}|\mathsf{X}_{i_1}\mathsf{Y}}(B_{i_n})...\mathbb{K}_y^{\mathsf{X}_{i_1}|\mathsf{Y}}(dx_{i_1}) \tag{2.106}$$

We can then apply Lemma 2.0.20 recursively to $\mathbb{K}_{x_{i_1},y}^{\mathsf{X}_{\rho(\{2,..,n\})}|\mathsf{X}_{i_1}\mathsf{Y}}(B_{i_n})$ and so forth. □

**Existence of Disintegrations**

The existence of disintegrations of standard measurable probability spaces is well known.

**Theorem 2.0.22** (Disintegration existence - probability space)**.** *Given a probability measure $\mathbb{P} \in \Delta(\mathcal{X}\otimes\mathcal{Y})$, if $(F,\mathcal{F})$ is standard then a disintegration $\mathbb{P}^{\mathsf{Y}|\mathsf{X}} : X \to \Delta(\mathcal{Y})$ exists (Çinlar, 2011).*

In particular, if for all $x \in X$, $\mathbb{P}^{\mathsf{X}}(\mathsf{X}\in\{x\}) > 0$, then $\mathbb{P}_x^{\mathsf{Y}|\mathsf{X}}(A) = \frac{\mathbb{P}^{\mathsf{XY}}(\{x\}\times A)}{\mathbb{P}^{\mathsf{X}}(\{x\})}$.

For Markov kernel spaces, standard measurability is not known to guarantee that a disintegration exists. Consider the following general setup: a kernel space $(\mathbb{K},(D,\mathcal{D}),(X\times Y,\mathcal{X}\otimes\mathcal{Y}))$ with $D, X$ and $Y$ all equal to $[0,1]$ and all Borel. Let $\mathsf{X},\mathsf{Y},\mathsf{D} : X\times Y\times D \to [0,1]$ project the first, second and third dimensions of $X\times Y\times D$ respectively. Let $\mathbb{K}_d(A) = \lambda(A)$, the Lebesgue measure of $A$ on $[0,1]^2$ for all $d\in D$.

By Theorem 2.0.22, we have for each $d\in D$ a disintegration $Q(d) := (\mathbb{K}_d)^{\mathsf{Y}|\mathsf{X}}$ of $(\mathbb{K}_d)^{\mathsf{XY}}$, and it is fairly straightforward to show it that $Q(d)_x(A) = \lambda(A)$ for all $A\in\mathcal{B}([0,1])$ and $\lambda$-almost all $x\in[0,1]$. $Q(d)_x$ is clearly a probability measure for every $(d,x)\in[0,1]^2$, but $Q : D\times X\to\Delta(\mathcal{Y})$ given by $(d,x,A)\mapsto Q(d)_x(A)$ may fail to be a Markov kernel.

To see this, let $\mathbb{1}_C : [0,1]\to\{0,1\}$ be the indicator function on a non-measurable set $C\subset[0,1]$, and define

$$Q(d)_x(A) = \left(1-\mathbb{1}_C(x)\mathbb{1}_{\{x\}}(d)\right)\lambda(A) + \mathbb{1}_C(x)\mathbb{1}_{\{x\}}(d)\delta_0(A) \tag{2.107}$$

That is, $Q$ is the measure $\lambda A$ for all points $(x,d)$ except where $x=d$ and $d\in C$. Note that for each value of $d$, $Q$ differs from $\lambda(A)$ on at most a single point $x\in[0,1]$, which has measure 0 under the Lebesgue measure $\lambda$. Thus $Q(d)$ is a version of $(\mathbb{K}_d)^{\{\mathsf{Y}|\mathsf{X}\}}$. Consider the function

$$Q^{\{0\}} : (d,x)\mapsto Q(d)_x(\{0\}) \tag{2.108}$$

$$\left(Q^{\{0\}}\right)^{-1}(\{1\}) = \{(d,x) : Q(d)_x(\{0\}) = 1\} \tag{2.109}$$

$$= \{(d,x) : d = x \ \ \& \ \ d\in C\} \tag{2.110}$$

Thus $Q^{\{0\}}$ is not measurable and consequently $Q$ fails to be a Markov kernel. The problem comes from the fact that $Q$ is defined by an uncountable collection of disintegrations $Q(d)$, each of which is individually measurable. In this case, the problem can be easily solved by defining $Q'$ without the non-measurable component in 2.107. What we would like are general conditions under which we know that we can choose an appropriate set of disintegrations $Q(d)$ in order for the resulting $Q$ to be a Markov kernel.

This problem can be easily dealt with if we only require $\mathbb{K}$ to be unique up to a set of measure 0 with respect to some "background probability" $\mathbb{P}^* \in \Delta(\mathcal{D} \otimes \mathcal{E})$, because we can simply take $\mathbb{K}$ to be an arbitrary disintegration of $\mathbb{P}^*$ and then use Theorem 2.0.20 to find further disintegrations (see Lemma 2.0.30. However, many common examples of causal models do admit a background probability. For example, with Causal Bayesian Networks, $do(\mathsf{X} = x)$ interventions are typically associated with a point probability measure $\delta_x$ on the intervened variable. If, for example, $X$ takes values in $[0, 1]$ then there is no probability measure that assigns nonzero probability to every real number we can choose for a do-intervention $do(\mathsf{X} = x)$.

For a more specific example with Causal Bayesian Networks, suppose we have $\mathsf{X}$ and $\mathsf{Y}$ in $[0, 1]$, the graph $\mathcal{G} := \mathsf{X} \to \mathsf{Y}$ and, as above, the observational distribution is $\mathbb{P}(\mathsf{X} \in A, \mathsf{Y} \in B) = \lambda(A)\lambda(B)$. Then, because elements of $\mathbb{P}^{\{\mathsf{Y}|\mathsf{X}\}}$ are unique up to a set of $\mathbb{P}$-measure 0, Equation 2.107 is a version of $\mathbb{P}^{\{\mathsf{Y}|\mathsf{X}\}}$ for each $d \in D$. Identify each $d \in D$ with an intervention $do(\mathsf{X} = x)$. According to the definition of Pearl (2009) pg 24, the interventional distribution $\mathbb{P}(d)(\mathsf{X}, \mathsf{Y})$ must have the following properties for evey $x \in X$:

- $\mathbb{P}^{\mathsf{XY}}_{do(\mathsf{X}=x)}$ is Markov relative to $\mathcal{G}$ (this condition is trivial with the given $\mathcal{G}$)
- $\mathbb{P}^{\mathsf{X}}_{do(\mathsf{X}=x)} = \delta_x$
- $\mathbb{P}^{\mathsf{XY}|\mathsf{X}}_{do(\mathsf{X}=x)}$ " $=''$ $\mathbb{P}^{\mathsf{XY}|\mathsf{X}}$, $\mathbb{P}_{do(\mathsf{X}=x)}$-almost surely

The quotation marks have been added to the final condition, as there are at least two different ways to interpret it:

- $\mathbb{P}^{\mathsf{XY}|\mathsf{X}}_{do(\mathsf{X}=x)} \in \mathbb{P}^{\{\mathsf{XY}|\mathsf{X}\}}$, $\mathbb{P}_{do(\mathsf{X}=x)}$-almost surely
- Let $\mathbb{T}^{\mathsf{XY}|\mathsf{X}}$ be a particular version of $\mathbb{P}^{\{\mathsf{XY}|\mathsf{X}\}}$. Then $\mathbb{P}^{\mathsf{XY}|\mathsf{X}}_{do(\mathsf{X}=x)} = \mathbb{T}^{\{\mathsf{XY}|\mathsf{X}\}}$, $\mathbb{P}_{do(\mathsf{X}=x)}$-almost surely

In the first case, we can choose an arbitrary element of $\mathbb{P}^{\{\mathsf{XY}|\mathsf{X}\}}$ for each $x \in X$. Furthermore, because each $\{x\} \in \mathcal{X}$ has $\mathbb{P}$-measure 0 but $\mathbb{P}_{do(\mathsf{X}=x)}$-measure 1, it is easy to verify that the third condition is satisfied for

$$\mathbb{P}^{\mathsf{XY}|\mathsf{X}}_{(\mathsf{X}=x),x'}(A \times B) = \mathbb{1}_C(x)\delta_x(A)\delta_1(B) + (1 - \mathbb{1}_C(x))\delta_x(A)\delta_0(B) \qquad (2.111)$$

As it is by any function at all $X \times X \to \Delta(\mathcal{X} \otimes \mathcal{Y})$. In this example we have for every $x \in X$, $\mathsf{Y}$ is with probability 1 an indicator of membership of $x$ in the non-measurable set $C$. This is a nonsensical result, depsite the apparent simplicity of the original causal model.

The second at least guarantees that $do(\mathsf{X} = x) \mapsto \mathbb{P}_{do\mathsf{X}=x}$ is measurable. However, it still allows for nonsense results. Letting $R$ be the Cantor set which has an uncountable number of elements and $\lambda(R) = 0$, it is still consistent to define

$$\mathbb{P}^{\mathsf{XY}|\mathsf{X}}_{(\mathsf{X}=x),x'}(A \times B) = \mathbb{1}_R(x)\delta_{x'}(A)\delta_1(B) + (1 - \mathbb{1}_R(x))\delta_x(A)\lambda(B) \tag{2.112}$$

While Equation 2.112 behaves "as it is supposed to" for mixtures of *do* operations $\pi \in \Delta([0,1])$ absolutely continuous with respect to the Lebesgue measure $\lambda \gg \pi$, it also sets $\mathsf{Y}$ to 1 for any point interventions in the Cantor set, or any mixtures of *do* operations absolutely continuous with respect to the Cantor set. This is possible because, in general, we don't have a rule for choosing a version of $\mathbb{P}^{\{\mathsf{XY}|\mathsf{X}\}}$ that assigns "sensible" values to all $\mathbb{P}$-measure 0 sets.

It is straightforward to show that wherever $D$ is countable, arbitrary disintegrations of $(\mathbb{K}, (D, \mathcal{D}), (E, \mathcal{E}))$ exist. **This is an assumption we will typically make.**

The following theorem establishes an alternative sufficient condition for the existence of disintegrations in a Markov kernel space. We introduce the notion of *ratio continuity* and show that a kernel that is ratio continuous or a countable piecewise combination of ratio continuous kernels that disintegrations exist. This implies the existence of disintegrations wherever $D$ is countable.

**Definition 2.0.23** (Ratio continuity)**.** A Markov kernel $\mathbb{K} : D \to \Delta(\mathcal{E})$ where $D$ is equipped with metric $m_D$ is *ratio continuious* if for every $d \in D$ and $\epsilon > 0$ there exists a $\delta > 0$ such that for all $A \in \mathcal{E}$ $m_D(d, d') < \delta \implies 1 - \epsilon \leq \frac{\mathbb{K}_d(A)}{\mathbb{K}_{d'}(A)} \leq \epsilon$.

This is a very strong notion of continuity - it implies that any set $A$ has $\mathbb{K}_d$-measure 0 if and only if it has $\mathbb{K}_{d'}$-measure 0 for all $d' \in D$.

**Theorem 2.0.24** (Existence of disintegrations on kernel spaces: uniform normalised continuous kernel)**.** *Given a kernel space $(\mathbb{K}, (D, \mathcal{D}), (E, \mathcal{E}))$ with $(D, \mathcal{D})$, $(E, \mathcal{E})$ standard measurable and some random variables $\mathsf{X} : E \times D \to X$, and $\mathsf{Y} : E \times D \to Y$, if for all $A \in \mathcal{E} \otimes \mathcal{D}$ the maps*

$$d \mapsto \mathbb{K}_d(A) \tag{2.113}$$

*are ratio continuous then for any $\mathsf{Y} : E \times D \to Y$ the disintegration $\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}$ exists.*

*Proof.* By standard measurability, $X$ and $Y$ are separable. In particular, there is some sequence $H_i \subset X$ such that $\sigma(\{H_i | i \in \mathbb{N}\}) = \mathcal{X}$. Then $\sigma(\{H_i | i \in [n]\})$ is finite and there exists some partition $\mathcal{J}_n \subset \sigma(\{H_i | i \in [n]\})$ such that $\sigma(\mathcal{J}_n) = \mathcal{X}$. Note that $\{\sigma(\mathcal{J}_n) | n \in \mathbb{N}\}$ is a filtration.

For each $x \in X$ and $n \in \mathbb{N}$, there is a unique $J_i \in \mathcal{J}_n$ such that $x \in J_i$. Take arbitrary $A \in \mathcal{Y}$, $d \in D$ and define

$$R_d^{A,n}(x) = \sum_{H_i \in \mathcal{J}_n} \mathbb{1}_{H_i}(x) \frac{\mathbb{K}_d^{\mathsf{XY}}(H_i \times A)}{\mathbb{K}_d^{\mathsf{XY}}(H_i \times Y)} \tag{2.114}$$

defining $\frac{0}{0} = 0$. Each $R_n^A$ is positive, $\sigma(\mathcal{J}_n)$-measurable and, because $\mathbb{P} \gg \mathbb{K}_d^{\mathsf{XY}}$,

$$\begin{aligned} \mathbb{E}[R_d^{A,n}(x)] &= \sum_i \mathbb{K}_d^{\mathsf{XY}}(H_i \times A) && (2.115)\\ &= \mathbb{K}_d^{\mathsf{XY}}(X \times A) && (2.116)\\ &< \infty && (2.117) \end{aligned}$$

Finally, taking $H_j \in \mathcal{J}_n$ and $H_i^{n+1} \in \mathcal{J}_{n+1}$:

$$\begin{aligned} \mathbb{E}[\mathbb{1}_{H_j} R_d^{A,n}] &= \int \mathbb{1}_{H_j}(x) \sum_i \mathbb{1}_{H_i}(x) \frac{\mathbb{K}_d^{\mathsf{XY}}(H_i \times A)}{\mathbb{K}_d^{\mathsf{XY}}(H_i \times Y)} d\mathbb{K}_d^{\mathsf{XY}}(x,y) && (2.118)\\ &= \mathbb{K}_d^{\mathsf{XY}}(H_j \times A) && (2.119)\\ &= \int \mathbb{1}_{H_j}(x) \sum_i \mathbb{1}_{H_i^{n+1}}(x) \frac{\mathbb{K}_d^{\mathsf{XY}}(H_i^{n+1} \times A)}{\mathbb{K}_d^{\mathsf{XY}}(H_i^{n+1} \times Y)} d\mathbb{K}_d^{\mathsf{XY}}(x,y) && (2.120)\\ &= \mathbb{E}[\mathbb{1}_{H_j} R_d^{A,n+1}] && (2.121) \end{aligned}$$

thus $\mathbb{E}[R_d^{A,n+1} | \mathcal{D}_n] = R_d^{A,n+1}$, so the sequence $\{R_d^{A,n} | n \in \mathbb{N}\}$ is a positive martingale. Furthermore, it is uniformly integrable (Lemma 2.0.27), so it converges to a measurable function $R_d^A$ almost surely and also in $L^1$.

For $H_{\in} \mathcal{X}$

$$\begin{aligned} \int_B R_d^A(x) d\mathbb{K}_d^{\mathsf{XY}}(\{x\} \otimes Y) = \lim_{n \to \infty} \int_B R_d^{A,n}(x) d\mathbb{K}_d^{\mathsf{XY}}(\{x\} \otimes Y) && (2.122)\\ && (2.123) \end{aligned}$$

By Equation 2.119, $\int_{H_j} R_d^A(x) d\mathbb{K}_d^{\mathsf{XY}}(\{x\} \otimes Y) = \mathbb{K}_d^{\mathsf{XY}}(H_j \times A)$ for all $H_j \in \cup n \in \mathbb{N} \mathcal{J}_n$. However, $\cup n \in \mathbb{N} \mathcal{J}_n$ is a p-system for $\mathcal{X}$ and so $\int_B R_d^A(x) d\mathbb{K}_d^{\mathsf{XY}}(\{x\} \otimes Y) = \mathbb{K}_d^{\mathsf{XY}}(B \times A)$ for all $B \in \mathcal{X}$.

Suppose $D = [0,1]$. This will later be generalised to a general standard measurable space using the isomorphism between $[0,1]$ and any uncountable standard measurable space.

Let $H^n(x)$ be $H_i \in \mathcal{D}_n$ such that $x \in H_i$. By ratio continuity of $\mathbb{K}$, we can choose for every $d$ and every $\epsilon > 0$ some $\delta$ such that $|d - d'| < \delta$ implies:

$$|R_d^{A,n}(x) - R_{d'}^{A,n}(x)| \leq \left| \frac{(1+\epsilon)\mathbb{K}_d^{\mathsf{XY}}(H(x)\times A) - \mathbb{K}_d^{\mathsf{XY}}(H(x)\times A)(1-\epsilon)}{\mathbb{K}_d^{\mathsf{XY}}(H(x)\times Y)(1+\epsilon)} \right| \tag{2.124}$$

$$= \left| \frac{2\epsilon\mathbb{K}_d^{\mathsf{XY}}(H(x)\times A)}{\mathbb{K}_d^{\mathsf{XY}}(H(x)\times Y)(1+\epsilon)} \right| \tag{2.125}$$

$$< 2\epsilon \tag{2.126}$$

For all $n \in \mathbb{N}$ and all $x \in X$.

Because $R_d^{A,n}$ converges almost surely to $R_d^n$ and $R_{d'}^{A,n}$ converges almost surely to $R_{d'}^{A,n}$, for $x$ such that $R_d^{A,n}$ and $R_{d'}^{A,n}$ both converge we have

$$|R_d^A(x) - R_{d'}^A(x)| \leq \inf_n \left| R_d^A(x) - R_d^{A,n}(x) + R_d^{A,n}(x) - R_{d'}^{A,n}(x) + R_{d'}^{A,n} - R_{d'}^A \right| \tag{2.127}$$

$$\leq \inf_n \left( |R_d^A(x) - R_d^{A,n}(x)| + |R_d^{A,n}(x) - R_{d'}^{A,n}(x)| + |R_{d'}^{A,n} - R_{d'}^A| \right) \tag{2.128}$$

$$\leq 2\epsilon \tag{2.129}$$

Thus for any $\epsilon > 0$ there is some $\delta > 0$ such $|d-d'| < \delta$ implies $|R_d^{A,n} - R_{d'}^{A,n}| \leq \epsilon$ except on some set $O_d \cup O_{d'}$ where $O_d$ is a set of $\mathbb{K}_d^{\mathsf{X}}$ measure 0 and $O_{d'}$ is a set of $\mathbb{K}_{d'}^{\mathsf{X}}$ measure 0. Note that $\mathbb{K}_{d'}^{\mathsf{X}}(O_d) = |\mathbb{K}_{d'}^{\mathsf{X}}(O_{d'}) - \mathbb{K}_d^{\mathsf{X}}(O_d)| \leq 0$ hence $\mathbb{K}_{d'}^{\mathsf{X}}(O_d) = 0$ and vise versa.

Let $\mathbb{Q}_D$ be the rationals between 0 and 1 and for each $r \in Q$ let $O_r$ be the set on which $R_d^{A,n}$ fails to converge, noting that $O := \cup_{r\in F} O_r$ is of $\mathbb{K}_d^{\mathsf{X}}$-measure 0 for all $d \in D$.

Choose some $y_0 \in Y$ and define for arbitrary $A \in \mathcal{Y}$

$$S^{A,n|\mathsf{XD}} := (x,d) \mapsto \mathbb{1}_{X\setminus O}(x) \sum_i^n \mathbb{1}_{\left[\frac{di}{n}, \frac{di+1}{n}\right]}(d) R_{\frac{di}{i}}^A(x) + \mathbb{1}_O(x)\delta_{y_0}(A) \tag{2.130}$$

where $\mathbb{P}$ is an arbitrary element of $\Delta(\mathcal{Y})$. Each $S^{A,n|\mathsf{XD}}$ is measurable because it is a sum of measurable functions. Furthermore, it is clear that if $x \in X \setminus O$, $S^{A,n|\mathsf{XD}}(d,x) \to R_d^A(x)$ as $n \to \infty$ and on $x \in O$, $S^{A,n|\mathsf{XD}}(d,x) \to \delta_{y_0}(A)$, and so the sequence $S^{A,n|\mathsf{XD}}$ goes to a limit:

$$S_d^{A|\mathsf{XD}}(x) := \mathbb{1}_{X\setminus O}(x) R_d^A(x) + \mathbb{1}_O(x)\delta_{y_0}(A) \tag{2.131}$$

Finally, note that for all $A \in \mathcal{Y}$, $B \in \mathcal{X}$

$$\int_B S_d^{A|\mathsf{XD}}(x)d\mathbb{K}_d^{\mathsf{X}}(x) = \int_B (\mathbb{1}_{X\setminus O}(x)R_d^A(x) + \mathbb{1}_O(x)\delta_{y_0}(A))d\mathbb{K}_d^{\mathsf{X}} \tag{2.132}$$

$$= \int_B R_d^A(x)d\mathbb{K}_d^{\mathsf{X}}(x) \tag{2.133}$$

$$= \mathbb{K}_d^{\mathsf{XY}}(B\times A) \tag{2.134}$$

Thus $S^{A|\mathsf{XD}} = \mathbb{E}[\mathbb{1}_A|\mathsf{XD}]$. All that remains to be shown is that $A \mapsto S_d^{A|\mathsf{XD}}(x)$ is a probability measure for all $x \in X$, $d \in D$. This is a standard argument that can be found, for example, in Çinlar (2011) pp. 151-152

which I'll add here next

□

Theorem 2.0.24 is made quite limiting by the requirement for ratio continuity - for example, it requires a kernel $\mathbb{K}_d$ where the measure 0 sets are the same for every $d \in D$. This can be relaxed somewhat by the fact that a countable set of such kernels can be combined piecewise and still yield a disintegrable kernel.

**Theorem 2.0.25** (Piecewise uniform normalized continuous kenrel)**.** *Given a kernel space $(\mathbb{K}, (D,\mathcal{D}), (E,\mathcal{E}))$ with $(D,\mathcal{D})$, $(E,\mathcal{E})$ standard measurable and some random variables $\mathsf{X} : E\times D \to X$, and $\mathsf{Y} : E\times D \to Y$ and a countable partition $\mathcal{J}$ of $X$, if there exists a set of kernels $\{\mathbb{K}^i | i \in \mathbb{N}\}$ such that for all $d \in D$, $B\times A \in \mathcal{X}\otimes\mathcal{Y}$*

$$\mathbb{K}_d^{\mathsf{XY}}(B\times A) = \sum_{J_i\in\mathcal{J}} \mathbb{1}_{J_i}(d)\mathbb{K}_d^{i,\mathsf{XY}}(B\times A) \tag{2.135}$$

*and each $\mathbb{K}^i$ is uniform normalized continuous on $J_i$ then the disintegration $\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}$ exists.*

*Proof.* By Theorem 2.0.24 we have for each $i$ a disintegration $\mathbb{K}^{i,\mathsf{Y}|\mathsf{XD}}$. Define

$$T_{x,d}^{\mathsf{Y}|\mathsf{XD}}(A) := \sum_{J_i\in\mathcal{J}} \mathbb{1}_{J_i}(d)\mathbb{K}_{x,d}^{i,\mathsf{Y}|\mathsf{XD}}(A) \tag{2.136}$$

We have

- $A \mapsto T_{x,d}^{\mathsf{Y}|\mathsf{XD}}(A)$ is a probability measure for each $x, d \in X\times D$ because this is true for each $\mathbb{K}^{i,\mathsf{Y}|\mathsf{XD}}$
- $(x,d) \mapsto T_{x,d}^{\mathsf{Y}|\mathsf{XD}}(A)$ is measurable for each $A \in \mathcal{Y}$ because it is a sum of measurable functions
- $(x,A) \mapsto T_{x,d}^{\mathsf{Y}|\mathsf{XD}}(A)$ is a version of $(\mathbb{K}_d)^{\mathsf{Y}|\mathsf{X}}$ for each $d \in D$ because this is also true for each $\mathbb{K}^{i,\mathsf{Y}|\mathsf{XD}}$

Thus $T^{\mathsf{Y}|\mathsf{XD}}$ is the required disintegration $\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}$. □

**Lemma 2.0.26.** *If $\mathbb{Q} \ll \mathbb{P}$ on $(E, \mathcal{E})$ then for all $\epsilon > 0$ there is some $\delta > 0$ such that for every $A \in \mathcal{E}$ $\mathbb{P}(A) < \epsilon \implies \mathbb{Q}(A) < \delta$*

*Proof.*

todo

□

**Lemma 2.0.27.** *$Q_d^{A,n}$ as define in Theorem 2.0.24 is uniformly integrable.*

*Proof.*

todo; a proof for an analagous fact is given in Çinlar (2011)

□

**Theorem 2.0.28** (Existence of disintegrations on kernel spaces: purely atomic measures)**.** *Given a kernel space $(\mathbb{K}, (D, \mathcal{D}), (\Omega, \mathcal{E}))$ with $(D, \mathcal{D})$ and $(\Omega, \mathcal{E})$ standard measurable, if $\mathbb{K}_d$ is purely atomic for all $d \in D$ then for any random variables $\mathsf{X}, \mathsf{Y} \in \mathcal{E} \otimes \mathcal{D}$ and domain variable $\mathsf{D} : \Omega \times D \mapsto D$ a disintegration $\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}$ exists.*

*Proof.*

show...

□

**Definition 2.0.29** (Relative probability space)**.**

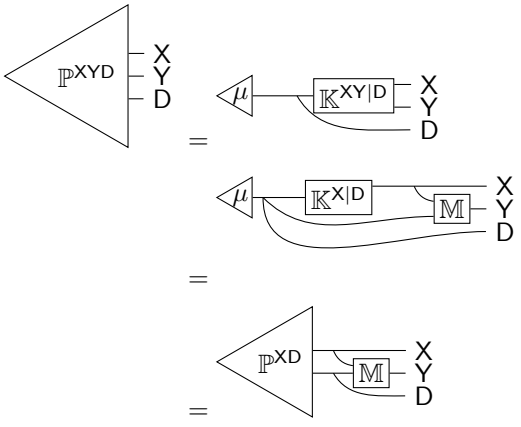
better name

Given a Markov kernel space $(\mathbb{K}, D, \Omega)$ and a strictly positive measure $\mu \in \Delta(\mathcal{D})$, $(\mu\mathbb{K}, \Omega \times D)$ is a *relative* probability space.

For any random variable $\mathsf{X} : \Omega \times D \to X$ on $(\mathbb{K}, D, \Omega)$, its relative on $(\mu\mathbb{K}, \Omega \times D)$ is given by the same measurable function, and we give it the same name $\mathsf{X}$.

**Lemma 2.0.30** (Agreement of disintegrations)**.** *Given a Markov kernel space $(\mathbb{K}, D, \Omega)$, any relative probability space $(\mu\mathbb{K}, \Omega \times D)$ and any random variables $\mathsf{X} : \Omega \times D \to X$, $\mathsf{Y} : \Omega \times D \to Y$, $\mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}} = (\mu\mathbb{K})^{\{\mathsf{Y}|\mathsf{XD}\}}$ (note that this set equality).*

*Proof.* Define $\mathbb{P} := \mu\mathbb{K}$ and let $\mathbb{M}$ be an arbitrary version of $\mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$. Then



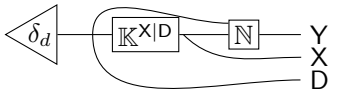

(2.137)

(2.138)

(2.139)

Thus $\mathbb{M} \in \mathbb{P}^{\{\mathsf{Y}|\mathsf{XD}\}}$.

Let $\mathbb{N}$ be an arbitrary version of $\mathbb{P}^{\{\mathsf{Y}|\mathsf{XD}\}}$. To show that $\mathbb{N} \in \mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$, we will show for all $d \in D$




$$\mathbb{Q} := \tag{2.140}$$

$$= \mathbb{K}_d^{\mathsf{XYD|D}} \tag{2.141}$$

For $A \in \mathcal{X}$,$B \in \mathcal{Y}$, $d \in D$, we have $\mathbb{Q}(A \times B \times \emptyset) = 0 = \mathbb{K}_d^{\mathsf{XYD|D}}(A \times B \times \emptyset$, and for $\{d\} \in \mathcal{D}$ we have $\mu(\{d\}) > 0$ so:

$$\mathbb{Q}(A \times B \times \{d\}) = \int_{X^2} \int_X \int_{D^3} \mathbb{N}_{d'',x'}(A)\mathbf{Id}_{x''}(B)\mathbf{Id}_{d'''}(\{d\}) d\curlyvee_d(d', d'', d''') d\mathbb{K}_{d'}^{\mathsf{X}|\mathsf{D}}(x) d\curlyvee_x(x', x'') \tag{2.142}$$

$$= \delta_d(\{d\}) \int_X \mathbb{N}_{d,x}(A)\delta_x(B) d\mathbb{K}_d^{\mathsf{X}|\mathsf{D}}(x) \tag{2.143}$$

$$= \frac{1}{\mu(\{d\})} \int_{\{d\}} d\mu(d') \int_X \mathbb{N}_{d,x}(A)\delta_x(B) d\mathbb{K}_d^{\mathsf{X}|\mathsf{D}}(x) \tag{2.144}$$

$$= \frac{1}{\mu(\{d\})} \int_D \int_X \mathbb{N}_{d,x}(A)\delta_{d'}(\{d\})\delta_x(B) d\mathbb{K}_d^{\mathsf{X}|\mathsf{D}}(a) d\mu(d') \tag{2.145}$$

$$= \frac{1}{\mu(\{d\})} \int_D \int_X \mathbb{N}_{d,x}(A)\delta_{d'}(\{d\})\delta_x(B) d\mathbb{K}_{d'}^{\mathsf{X}|\mathsf{D}}(a) d\mu(d') \tag{2.146}$$

$$= \frac{1}{\mu(\{d\})} \mathbb{P}^{\mathsf{XYD}}(A \times B \times \{d\}) \tag{2.147}$$

$$= \frac{1}{\mu(\{d\})} \int_D \mathbb{K}_{d'}^{\mathsf{XYD}|\mathsf{D}}(A \times B \times \{d\}) d\mu(d') \tag{2.148}$$

$$= \frac{1}{\mu(\{d\})} \int_D \mathbb{K}_{d'}\mathsf{XY}|\mathsf{D}(A \times B)\delta_{d'}(\{d\}) d\mu(d') \tag{2.149}$$

$$= \mathbb{K}_d^{\mathsf{XY}|\mathsf{D}}(A \times B) \tag{2.150}$$

$$= \mathbb{K}_d^{\mathsf{XY}|\mathsf{D}}(A \times B)\delta_d(\{d\}) \tag{2.151}$$

$$= \int_D \mathbb{K}_{d'}^{\mathsf{XY}}(A \times B)\delta_{d''}(\{d\}) d\curlyvee_d(d', d'') \tag{2.152}$$

$$= \mathbb{K}_d^{\mathsf{XYD}|\mathsf{D}}(A \times B \times \{d\}) \tag{2.153}$$

Equality follows from the monotone class theorem. Thus $\mathbb{N} \in \mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$. □

Thus any kernel conditional probability $\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}$ can equally well be considered a regular conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{XD}}$ for a related probability space $(\mathbb{P}, \Omega \times D)$ under the obvious identification of random variables, provided $D$ is countable. Note that any conditional probability $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ that is *not* conditioned on $\mathsf{D}$ is undefined in the kernel space $(\mathbb{K}, D, \Omega)$.

### Conditional Independence

**Definition 2.0.31** (Kernels constant in an argument)**.** Given a kernel $(\mathbb{K}, D, \Omega)$ and random variables $\mathsf{Y}$ and $\mathsf{X}$, we say a verstion of the disintegration $\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}$ is constant in $\mathsf{D}$ if for all $x \in X$, $d, d' \in D$, $\mathbb{K}_{(x,d)}^{\mathsf{Y}|\mathsf{XD}} = \mathbb{K}_{(x,d')}^{\mathsf{Y}|\mathsf{XD}}$.

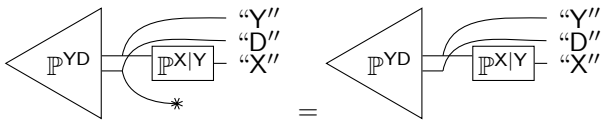
**Definition 2.0.32** (Domain Conditional Independence)**.** Given a kernel space $(\mathbb{K}, D, \Omega)$, relative probability space $(\mathbb{P}, \Omega \times D)$, variables $\mathsf{X}$,$\mathsf{Y}$ and domain

variable $\mathsf{D}$, $\mathsf{X}$ is *conditionally independent* of $\mathsf{D}$ given $\mathsf{Y}$, written $\mathsf{X} \perp\!\!\!\perp_{\mathbb{K}} \mathsf{D}|\mathsf{Y}$ if any of the following equivalent conditions hold:

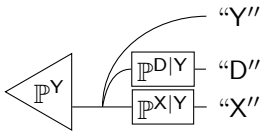
Almost sure equality

1. $\mathbb{P}^{\mathsf{XD}|\mathsf{Y}} \sim \mathbb{P}^{\mathsf{X}|\mathsf{Y}} \underline{\otimes} \mathbb{P}^{\mathsf{D}|\mathsf{Y}}$
2. For any version of $\mathbb{P}^{\{\mathsf{X}|\mathsf{Y}\}}$, $\mathbb{P}^{\mathsf{X}|\mathsf{Y}} \otimes *_D$ is a version of $\mathbb{K}^{\{\mathsf{X}|\mathsf{YD}\}}$
3. There exists a version of $\mathbb{K}^{\{\mathsf{X}|\mathsf{YD}\}}$ constant in $\mathsf{D}$

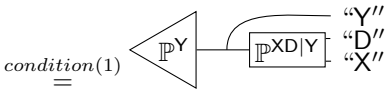
**Theorem 2.0.33** (Definitions are equivalent)**.** *(1) $\implies$ (2): By Lemma 2.0.30, $\mathbb{P}^{\{\mathsf{Y}|\mathsf{XD}\}} = \mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$. Thus it is sufficient to show that $\mathbb{P}^{\mathsf{X}|\mathsf{Y}} \otimes *$ is a version of $\mathbb{P}^{\{\mathsf{X}|\mathsf{YD}\}}$.*



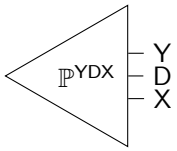

(2.154)




(2.155)




(2.156)



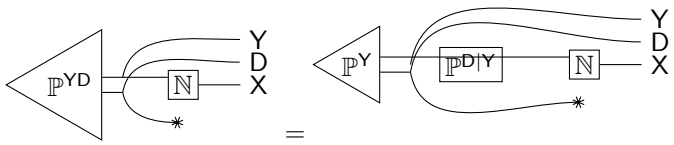

(2.157)

*(2) $\implies$ (3)*

*$\mathbb{P}^{\mathsf{X}|\mathsf{Y}} \otimes *_D$ is a version of $\mathbb{K}^{\{\mathsf{X}|\mathsf{YD}\}}$ by assumption, and is clearly constant in* $\mathsf{D}$.

*(3) $\implies$ (1)*

*By lemma 2.0.30, there also exists a version of $\mathbb{P}^{\{\mathsf{X}|\mathsf{YD}\}}$ constant in* $\mathsf{D}$*. Let $\mathbb{M} : Y \times D \to \Delta(\mathcal{X})$ be such a version. For arbitrary $d_0 \in D$, let $\mathbb{N} := \mathbb{M}_{(\cdot,d_0)} : Y \to \Delta(\mathcal{X})$ be the map $x \mapsto \mathbb{M}_{(x,d_0)}$. By constancy in* $\mathsf{D}$*, $\mathbb{M} = * \otimes \mathbb{N}$. We wish to show $\mathbb{P}^{\mathsf{X}|\mathsf{Y}} \underline{\otimes} \mathbb{P}^{\mathsf{D}|\mathsf{Y}} \in \mathbb{P}^{\{\mathsf{XD}|\mathsf{Y}\}}$. By Theorem 2.0.20, we have*




(2.158)

**Definition 2.0.34** (Conditional probability existence)**.** Given a kernel space $(\mathbb{K}, D, \Omega)$ and random variables $\mathsf{X}$, $\mathsf{Y}$, we say $\mathbb{K}^{\{\mathsf{Y}|\mathsf{X}\}}$ *exists* if $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{K}} \mathsf{D}|\mathsf{X}$. If $\mathbb{K}^{\{\mathsf{Y}|\mathsf{X}\}}$ exists then it is by definition equal to $\mathbb{P}^{\{\mathsf{Y}|\mathsf{X}\}}$ for any related probability space $(\mathbb{P}, \Omega \times D)$.

Note that $\mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$ always exists.

**Definition 2.0.35** (Conditional Independence)**.** Given a kernel space $(\mathbb{K}, D, \Omega)$, some relative probability space $(\mathbb{P}, \Omega \times D)$, variables $\mathsf{X}$,$\mathsf{Y}$ and $\mathsf{Z}$, $\mathsf{X}$ is *conditionally independent* of $\mathsf{Z}$ given $\mathsf{Y}$, written $\mathsf{X} \perp\!\!\!\perp_{\mathbb{K}} \mathsf{Z}|\mathsf{Y}$ if $\mathbb{K}^{\{\mathsf{XY}|\mathsf{Z}\}}$ exists and any of the following equivalent conditions hold:

Almost sure equality

- $\mathbb{P}^{\mathsf{XZ}|\mathsf{Y}} \sim \mathbb{P}^{\mathsf{X}|\mathsf{Y}} \underline{\otimes} \mathbb{P}^{\mathsf{Z}|\mathsf{Y}}$
- For any version of $\mathbb{P}^{\{\mathsf{X}|\mathsf{Y}\}}$, $\mathbb{P}^{\mathsf{X}|\mathsf{Y}} \otimes *_Z$ is a version of $\mathbb{K}^{\{\mathsf{X}|\mathsf{YZ}\}}$
- There exists a version of $\mathbb{K}^{\{\mathsf{X}|\mathsf{YZ}\}}$ constant in $\mathsf{Z}$

**Lemma 2.0.36** (Diagrammatic consequences of labels)**.** *In general, diagram labels are "well behaved" with regard to the application of any of the special Markov kernels: identities 2.17, swaps 2.28, discards 2.34 and copies 2.25 as well as with respect to the coherence theorem of the CD category. They are not "well behaved" with respect to composition.*

*Fix some Markov kernel space $(\mathbb{K}, D, \Omega)$ and random variables $\mathsf{X}$, $\mathsf{Y}$, $\mathsf{Z}$ taking values in $X, Y, Z$ respectively.* Sat : *indicates that a labeled diagram satisfies definitions 2.0.15 and 2.0.18 with respect to $(\mathcal{K}, D, \Omega)$ and $\mathsf{X}$, $\mathsf{Y}$, $\mathsf{Z}$. The following always holds:*

$$\text{Sat} : \mathsf{X} \text{ — } \mathsf{X} \tag{2.159}$$

*and the following implications hold:*

$$\text{Sat} : \mathsf{Z} \text{ — } \boxed{\mathbb{K}} \text{ — } \begin{matrix}\mathsf{X}\\ \mathsf{Y}\end{matrix} \implies \text{Sat} : \mathsf{Z} \text{ — } \boxed{\mathbb{K}} \text{ — } \mathsf{X} \quad * \tag{2.160}$$

$$\text{Sat} : \mathsf{Z} \text{ — } \boxed{\mathbb{K}} \text{ — } \begin{matrix}\mathsf{X}\\ \mathsf{Y}\end{matrix} \implies \text{Sat} : \mathsf{Z} \text{ — } \boxed{\mathbb{K}} \text{ — } \begin{matrix}\mathsf{Y}\\ \mathsf{X}\end{matrix} \tag{2.161}$$

$$\text{Sat} : \mathsf{Z} \text{ — } \boxed{\mathbb{L}} \text{ — } \mathsf{X} \implies \text{Sat} : \mathsf{Z} \text{ — } \boxed{\mathbb{L}} \text{ — } \begin{matrix}\mathsf{X}\\ \mathsf{X}\end{matrix} \tag{2.162}$$

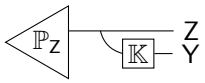
$$\text{Sat} : \mathsf{Z} \text{ — } \boxed{\mathbb{K}} \text{ — } \mathsf{Y} \implies \text{Sat} : \mathsf{Z} \text{ — } \begin{matrix}\mathsf{Z}\\ \boxed{\mathbb{K}} \text{ — } \mathsf{Y}\end{matrix} \tag{2.163}$$

*Proof.*

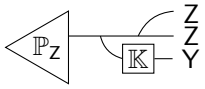
- $\mathrm{Id}_X$ is a version of $\mathbb{P}_{\mathsf{X}|\mathsf{X}}$ for all $\mathbb{P}$; $\mathbb{P}_{\mathsf{X}}\mathrm{Id}_X = \mathbb{P}_{\mathsf{X}}$
- $\mathbb{K}\mathrm{Id} \otimes *)(w; A) = \int_{X\times Y} \delta_x(A)\mathbb{1}_Y(y)d\mathbb{K}_w(x, y) = \mathbb{K}_w(A \times Y) = \mathbb{P}_{\mathsf{X}|\mathsf{Z}}(w; A)$

- $\int_{X\times Y} \delta_{\mathrm{swap(x,y)}}(A\times B)d\mathbb{K}_w(x,y) = \mathbb{P}_{\mathsf{YX|Z}}(w; A\times B)$
- $\mathbb{K}\curlyvee(w; A\times B) = \int_X \delta_{x,x}(A\times B)d\mathbb{K}_w(x) = \mathbb{P}_{\mathsf{XX|Z}}(w; A\times B)$

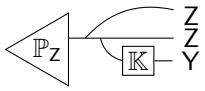
2.163: Suppose $\mathbb{K}$ is a version of $\mathbb{P}_{\mathsf{Y|Z}}$. Then

$\mathbb{P}_{\mathsf{ZY}} =$  (2.164)



$\mathbb{P}_{\mathsf{ZZY}} =$  (2.165)



$=$  (2.166)



Therefore $\curlyvee(\mathrm{Id}_X \otimes \mathbb{K})$ is a version of $\mathbb{P}_{\mathsf{ZY|Z}}$ by Definition 2.0.34 □

The following property, on the other hand, does *not* generally hold:

$$\text{Sat} : \mathsf{Z} \text{ —}\boxed{\mathbb{K}}\text{— } \mathsf{Y} \text{ , } \mathsf{Y} \text{ —}\boxed{\mathbb{L}}\text{— } \mathsf{X} \implies \text{Sat} : \mathsf{Z} \text{ —}\boxed{\mathbb{K}}\text{—}\boxed{\mathbb{L}}\text{— } \mathsf{X} \tag{2.167}$$

Consider some ambient measure $\mathbb{P}$ with $\mathsf{Z} = \mathsf{X}$ and $\mathbb{P}_{\mathsf{Y|X}} = x \mapsto \text{Bernoulli}(0.5)$ for all $z \in Z$. Then $\mathbb{P}_{\mathsf{Z|Y}} = y \mapsto \mathbb{P}_{\mathsf{Z}}$, $\forall y \in Y$ and therefore $\mathbb{P}_{\mathsf{Y|Z}}\mathbb{P}_{\mathsf{Z|Y}} = x \mapsto \mathbb{P}_{\mathsf{Z}}$ but $\mathbb{P}_{\mathsf{Z|X}} = x \mapsto \delta_x \neq \mathbb{P}_{\mathsf{Y|Z}}\mathbb{P}_{\mathsf{Z|Y}}$.

# Chapter 3

# Two player statistical models and see-do models

These are "todo" notes. All such notes that involve theoretical development are also collected in an unordered list of outstanding theoretical questions

In this chapter I introduce two types of model. Models of the first type are called *two player statistical models* and the second type are a special class of the first called *see-do models*. Fundamentally, each of these is just a particular kind of stochastic function. The reason we are interested in these kinds of stochastic functions is that is that almost all causal models are instances of see-do models. Before introducing two player models and discussing what makes them causal, it is worth briefly considering models in statistics and machine learning generally.

A *world model* is something I will informally define as a family of "descriptions" indexed by hypotheses $\{R_h | h \in H\}$. The set $H$ represents hypotheses or proposals for how the world ought to be described, and each proposal $h \in H$ entails some description of the world $\mathbb{R}_h$. Some examples of world models:

- A linear regressor may take some data $\mathbf{x}$ and $\mathbf{y}$ and returns a parameter $\beta \in B$ with the property that $(\mathbf{y} - \mathbf{x}^T\beta)^2$ is small. A normal way to interpret the parameter $\beta$ is to consider it to be a proposal about how some phenomenon of interest should be described, with this description explicitly given by the function $f : x \mapsto \beta x$.

- A neural network used in classification may take data $\mathbf{x}$ and labels $\mathbf{y}$ and returns parameters $\mathbf{w} \in W$ with the property that $-\mathbf{y}\log\mathbf{x} + (1 - y)\log(1 - \mathbf{x})$ is small. Each $\mathbf{w}$ is a proposal for how to classify data and the classification rule associated with each $\mathbf{w}$ is a function $x \mapsto f(\mathbf{w}, x)$.

- A crude description of a general election pre-poll result can be given by the "true fraction" $\theta$ of voters for each candidate and, under some unreasonably strong sampling assumptions, and the results of the survey for each $\theta$ can

be described by $\prod_N \mathbb{P}_\theta^{\mathsf{X}}$ where $N$ is the number of voters surveyed and $\mathsf{X}$ is the vote choice of each.

In the first two examples the "description" that goes with each hypothesis is a function, while in the third example the descriptions are probability measures. In almost all practical cases, these descriptions of the world do not tell us exactly how the world will turn out under each hypothesis, but at best offer us a prediction that is as good as we can hope for. Probability is the tool that is very widely used to formalise such "descriptions with uncertainty". Say I have two different linear regressors: one which minimises squared error on the training data and one that always returns $\beta = 10$. I want to ask which one produces descriptions that are more fit for my purpose. It is pointless to ask which one is correct because, in general, I cannot know that either will offer a description that is even approximately correct. However, I can consider a second level world model $\{\mathbb{P}_\alpha^{\mathsf{XY}} | \alpha \in A\}$ in which the phenomenon of interest is described by a family of probability measures, and then I can ask, given an $\alpha$, which $\beta$ is my regressor likely to return and how closely will $x \mapsto \beta x$ be to $\mathbb{E}_{\mathbb{P}_\alpha}[\mathsf{Y}|\mathsf{X}]$ for each likely choice. Generally, if I need to model a world with uncertainty I will need a world model that is an indexed family of probability measures.

A world model that consists of a family of probability measures $\{\mathbb{P}_h | h \in H\}$ is a *statistical model* or *statistical experiment*. Because I almost always need to Statistical models can be found everywhere in theoretical statistics and machine learning Fisher (1992); Le Cam (1996); Freedman (1963); de Finetti (1992); Vapnik (2013); Wald (1950). A key point about statistical models – even if I can only state it somewhat vaguely – is that the truth of any hypothesis $h \in H$ has no dependence on what I might want to be true. As a user of statistical models, I have no authority to choose a hypothesis – this is Nature's choice alone.

I can sometimes make choices that will affect the way that the future turns out. I might have some set $D$ of choices I can make, and for each $d \in D$ I require a description of the results of my choice. Just as the results of hypotheses are often uncertain, so are the results of choices. I might be motivated to choose a probability measure $\mathbb{P}_d$ to describe them, maybe because it is common to do so or because I find arguments for subjective expected utility theory compelling (Steele and Stefánsson, 2020). A family of probability measures indexed by a set of choices $\{\mathbb{P}_d | d \in D\}$ will be called a *consequence model*.

Statistical models and consequence models are both families of probability measures indexed by arbitrary sets, which we have called hypotheses $H$ and choices $D$ respectively. These sets are distinguished by how they are interpreted when a two-player statistical model is used in the course of solving some kind of problem. The difference can be informally summarised in this manner: I do not get to tell Nature what choice $h \in H$ she makes, and Nature does not get to tell me what choice $d \in D$ I make. It will often be the case that I have multiple choices that can affect how the world turns out *and* I have multiple hypotheses about how each choice will affect the world. In this case, I will have a *two-player statistical model* $\{\mathbb{P}_{h,d} | h \in H, d \in D\}$.

So far I have explained the distinction between "player 1" and "player 2"

in vague metaphorical terms. If I am using a two-player statistical model in the context of a well defined problem such as "given data, what choice should I make?" then we can say precisely what $H$ and $D$ are and what role each plays in the problem. However, the field of causal inference includes other types of problem such as counterfactual problems which involve a choice set $D$ that plays a different role to the choice set in decision problems. Thus, while I will argue that causal models are two-player statistical models, and the second player is what distinguishes them from ordinary statistical models, the same kind of model can be used with different interpretations of what the second player's choices represent. This will be explored in more detail in the coming chapters.

Note to proof readers: I moved the discussion of decomposability to the next chapter so I can introduce it alongside the result that uses it

## 3.1 Two player statistical models and see-do models

Two player statistical models were introduced as doubly indexed sets of probability measures $\{\mathbb{P}_{h,d} | h \in H, d \in D\}$. If each $\mathbb{P}_{h,d} \in \Delta(\mathcal{E})$ for some measurable space $(E, \mathcal{E})$, the indexed set is equivalent to a function $H \times D \to \Delta(\mathcal{E})$. In the following work, we will make two simplifying assumptions:

1. A two player statistical model can be represented by a *Markov kernel* $\mathbb{T} : H \times D \to \Delta(\mathcal{E})$
2. The kernel space $(\mathbb{T}, (H \times D, \mathcal{H} \otimes \mathcal{D}), (E, \mathcal{E}))$ admits disintegrations $\mathbb{T}^{\mathsf{Y}|\mathsf{XDH}}$ for arbitrary random variables $\mathsf{X}, \mathsf{Y}$ on $H \times D \times E$ and domain variable $\mathsf{D} \underline{\otimes} \mathsf{H}$

The first condition amounts to the additional requirement that $(h, d) \mapsto \mathbb{T}_{h,d}(A)$ is measurable for every $A \in \mathcal{H} \otimes \mathcal{D} \otimes \mathcal{E}$, and sufficient for the second condition is that $D \times H$ is countable and $X \times Y$ standard measurable (though this is not necessary, see Theorem 2.0.24).

**Definition 3.1.1** (Two player statistical model)**.** A *two-player statistical model* $(\mathbb{T}, \mathsf{H}, \mathsf{D}, \mathsf{O})$ is a Markov kernel $\mathbb{T} : H \times D \to \Delta(\mathcal{O})$ such that, 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 along with three distinguished random variables: the *hypothesis* $\mathsf{H} : H \times D \times O \to H$ given by $(h, d, o) \mapsto h$ (forgetting the choice and outcome) and the *choice* $\mathsf{D} : H \times D \times O \to D$ given by $(h, d, o) \mapsto d$ (forgetting the hypothesis and outcome) and the *outcome* $\mathsf{O} : H \times D \times O \to O$ given by $(h, d, o) \mapsto o$ (forgetting the choice and hypothesis).

Decision problems involving often involve some data $\mathsf{X}$ is observed, then a choice is made, then the consequences $\mathsf{Y}$ are observed. In such a model, the observed data $\mathsf{X}$ cannot be affected by the choice. These models will be called *see-do* models to capture the assumption that there is an order in which seeing and doing happen.

**Definition 3.1.2** (See-Do model)**.** A *see-do model* $(\mathbb{T}, \mathsf{H}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ is a two-player statistical model $(\mathbb{T}, \mathsf{H}, \mathsf{D}, \mathsf{O})$ with two additional distinguished random variables: the *observation* $\mathsf{X} : H \times D \times O \to X$ and the *consequence* $\mathsf{Y} : H \times D \times O \to Y$ such that the outcome is the coupled product of the observation and the consequence $\mathsf{O} = \mathsf{X} \underline{\otimes} \mathsf{Y}$. A see-do model must observe the conditional independence $\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D} | \mathsf{H}$, i.e. the observation is independent of the choice conditional on the hypothesis.

Because $\mathsf{O} = \mathsf{X} \underline{\otimes} \mathsf{Y}$, we do not need to explicitly define $\mathsf{O}$ when specifying a see-do model.

## 3.2 Frequentist random variables and Bayesian forecasts

We've chosen to represent the "description of the world" using probability. This is an overwhelmingly common choice for describing things with uncertainty, but it is worth asking precisely what is being described here. It's well-known that probability is suitable for representing a number of different things. Two common choices are:

1. The long run convergence of relative frequencies of sequences or ensembles of observations of certain types of systems (*frequentist probability*)

2. Forecasts of observations that will take place in the future, or, more generally, forecasts of things which we are uncertain about for any reason (*Bayesian forecasts*)

The first view is a very common interpretation of probability as it is used in statistical models. One can view each hypotheses in a statistical model as representing the proposition that the system will tend to produce long sequences of observations with relative frequencies consistent with the associated probability measure. For example, given a possibly loaded die, we might entertain hypotheses a) it is a system that produces a 6 $\frac{1}{6}$ of the time, b) it is a system that produces a 6 $\frac{1}{4}$ of the time, and many other hypotheses besides.

On the other hand, if I view a sequence of random variables as a sequence of *Bayesian forecasts* then I do not strictly need to entertain a set of hypotheses regarding the long run relative frequencies of observations. If $\mathsf{X}_1, \mathsf{X}_2, \mathsf{X}_3, ...$ are rolls of a possibly loaded die, then I can make forecasts from a single probability distribution over $\mathsf{X}_{[n]}$ (sometimes called the "posterior distribution"). Furthermore, this distribution need not have any particular frequentist properties: I can forecast the probability of a 6 is, for each roll, $1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1$ with no dependence on previously seen rolls, such that relative frequencies never converge. Provided my forecast is a valid probability measure, it may seem unwise but it is still a well formulated Bayesian forecast.

Consider a see-do model in which $\mathsf{Y} \perp\!\!\!\perp \mathsf{D} | \mathsf{H}$ - that is, fixing each hypothesis, the consequence $\mathsf{Y}$ is independent of the choice $\mathsf{D}$. This may be a model of a situation in which we consider a number of hypotheses possible, and for

each hypothesis the consequences we will observe will actually be independent of the choices we make (where "actually independent" means something like "independent in the frequentist limit", acknowledging that this isn't a fully satisfactory definition, see for example Hájek (2019)). However, it could also model a situation in which we consider a number of hypotheses viable and within each hypothesis we are indifferent between different relationships $\mathsf{D}$ and $\mathsf{Y}$ might have, without expressing any view on whether the values obtained from repeated experiments would be independent in the frequentist sense (in any case, such repetitions may be impossible).

While forecasts are more commonly associated with a single probability measure and frequentist models with sets of probability measures indexed by hypotheses (also called parameters or states), this does not have to be the case. For example, Walley (1991) develops a theory of reasoning under uncertainty he calls *imprecise probability* that makes use of sets of probability measures.

Properties of see-do models like $\mathsf{Y} \perp\!\!\!\perp \mathsf{D}|\mathsf{H}$ may sometimes be objective claims about relationships that may or may not be observed under future experimentation. At the same time, the names given to the random variables – $\mathsf{Y}$ is the *consequence* and $\mathsf{D}$ is the *choice* – suggest that these variables may often refer to things that haven't happened yet. In fact, this is explicit in the following chapter where we consider using see-do models to make decisions – if I am considering which decision I should make, then by definition I have not made that decision yet. Variables that refer to things that haven't happened yet are, as mentioned above, forecasts.

So it seems that we sometimes want to view see-do models as Bayesian forecasts of the consequences of choices, and sometimes we also want properties of see-do models to describe "objective facts" of some sort. In Chapter 5 we will discuss how Causal Bayesian Networks are a particular kind of see-do model, so any discussion of what we want to use see-do models to represent also has some bearing on what we want to use Causal Bayesian Networks to represent. Furthermore, in the first chapter of Pearl (2009) we find examples of both of the views we have considered so far. Firstly, in introducing probability Pearl writes

> We will adhere to the Bayesian interpretation of probability, according to which probabilities encode degrees of belief about events in the world and data are used to strengthen, update, or weaken those degrees of belief. In this formalism, degrees of belief are assigned to propositions (sentences that take on true or false values) in some language, and those degrees of belief are combined and manipulated according to the rules of probability calculus.

While later he writes

> [...] causal relationships are ontological, describing objective physical constraints in our world, whereas probabilistic relationships are epistemic, reflecting what we know or believe about the world.

Causal Bayesian Networks are see-do models, so if Causal Bayesian Networks describe causal relationships and causal relationships are objective facts then

see-do models must describe objective facts. What exactly does it mean for see-do models to describe objective facts?

In the world of "ordinary" one-player statistical models, De Finetti's Representation Theorem establishes that if we are using probabilities to represent forecasts and we are willing to assume that these forecasts are *exchangeable* – that is, the probability distribution we choose to represent our forecast of future events is exactly the same as the probability distribution we would choose to represent our forecasts of any permutation of these events – then it is possible to define a random variable representing the *hypothesis* such that the probability distribution representing our forecast can be written as the product of a *prior distribution* over hypotheses and a statistical model that maps hypotheses to independent and identically distributed sequences of observations, which is a sequence of precisely the type that converges in the frequentist sense.

Thus, one way that statistical models could describe objective facts is if forecasts can be assumed to be exchangeable, in which case the forecast can be factored into a prior and a statistical model describing independent and identically distributed sequences. As Walley (1991) cautions, exchangeability is quite a strong assumption – for example, it entails that one will never believe, no matter how much evidence is apparent, that a sequence is cyclic rather than converging to a stable relative frequency. Acknowledging there is a lot more to explore in this area, in Theorem 3.2.13 we extend De Finetti's representation theorem to see-do models and show that, considering *see-do forecasts*, if both observations and consequences satisfy a kind of exchangeability then the forecast is the product of a prior and a see-do model which, given a hypothesis, describes a sequence of independent and identical consequence maps.

**Definition 3.2.1** (Forecasts, see-do forecast)**.** A *do forecast* $(\mathbb{F}, \mathsf{D}, \mathsf{O})$ is a Markov kernel $\mathbb{F} : D \to \Delta(\mathcal{O})$ for some set of choices $D$ and outcomes $O$. The choice variable $\mathsf{D} : D \times O \to D$ is the map $(d, e) \mapsto d$ that forgest the outcome and and the outcome variable $\mathsf{O} : D \times O \to O$ is the map $(d, o) \to o$ that forgets the choice.

A *see-do forecast* is a forecast $(\mathbb{F}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ with an *observation variable* $\mathsf{X} : D \times O \to X$ and a *consequence variable* $\mathsf{Y} : D \times O \to Y$ such that $\mathsf{O} = \mathsf{X} \underline{\otimes} \mathsf{Y}$ and $\mathsf{X} \perp\!\!\!\perp \mathsf{D}$.

A *forecast* $(\mathbb{F}, \mathsf{O})$ is a probability measure $\mathbb{F} \in \Delta(\mathcal{O})$ and an outcome variable $\mathsf{O} : O \to O$.

We can go from a see-do model to a see-do forecast by adding a prior to the model.

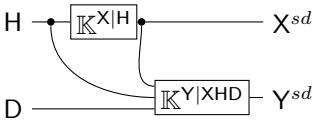
**Theorem 3.2.2.** *Given a two player statistical model* $(\mathbb{K}, \mathsf{H}, \mathsf{D}, \mathsf{O})$ *and any prior* $\mu \in \Delta(\mathcal{H})$*, defining* $\mathbb{L} := (\mu \otimes \mathrm{Id}_D)\mathbb{K}$ *we have* $(\mathbb{L}, \mathsf{D}, \mathsf{O})$ *is a do-forecast.*

*If* $(\mathbb{K}, \mathsf{H}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ *is a see-do model then, defining* $\mathbb{L}$ *as before,* $(\mathbb{L}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ *is a see-do forecast.*

*Proof.* The first part is trivial: $(\mu \otimes \mathrm{Id}_D)\mathbb{K}$ is a Markov kernel $D \to \Delta(\mathcal{O})$ by construction, and $\mathsf{D}$ and $\mathsf{O}$ are choice and outcome variables by definition of the original two player statistical model.

For the second part $\mathsf{X} \perp\!\!\!\perp_{\mathbb{L}} \mathsf{D}$ is required. By Theorem 2.0.20 we have

$$\mathbb{K} = \mathbb{K}^{\mathsf{XY}|\mathsf{HD}} \tag{3.1}$$

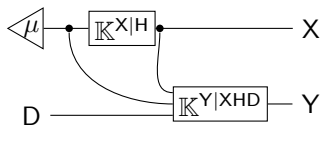



$$= \tag{3.2}$$

It follows that

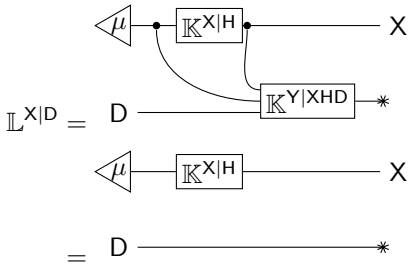
$$\mathbb{L}^{\mathsf{XY}|\mathsf{D}} = \mathbb{L} \tag{3.3}$$

$$= (\mu \otimes \mathrm{Id}_D)\mathbb{K} \tag{3.4}$$




$$= \tag{3.5}$$

Then




$$\tag{3.6}$$



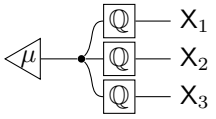

$$\tag{3.7}$$

And so $\mathsf{X} \perp\!\!\!\perp_{\mathbb{L}} \mathsf{D}$. □

In addition, any see-do forecast can be interpreted as a see-do model with a single hypothesis. Recalling the discussion of the indiscrete space $\{*\}$ in 2.0.3, we can identify a Markov kernel $\mathbb{F} : D \to \Delta(\mathcal{E})$ with a Markov kernel $\mathbb{T} : \{*\} \times D \to \Delta(\mathcal{E})$ where $\mathbb{T}_{*,d} = \mathbb{F}_d$ for all $d \in D$. Defining the hypothesis $\mathsf{H} : \{*\} \times D \times E \to H$ given by the constant fuction $(*, d, e) \mapsto *$, we can create from any see-do forecast $(\mathbb{F}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ a see-do model $(\mathsf{T}, \mathsf{H}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ (the required conditional independence is observed by construction in the single hypothesis $*$). However, this single hypothesis model is typically not a *frequentist model*.

de Finetti (1992) has shown how frequentist models in particular can be recovered from exchangeable forecasts. Informally speaking, if and only if a forecast $(\mathbb{P}, \mathsf{O})$ has the property that distribution of a sequence of random variables $\mathbb{P}^{\mathsf{X}_1\mathsf{X}_2\mathsf{X}_3}$ is identical to the distribution of any permutation of the sequence $\mathbb{P}^{\mathsf{X}_2\mathsf{X}_1\mathsf{X}_3}$ (an assumption known as *exchangeability*), and this sequence

can be extended infinitely, then there exists a hypothesis class $(H, \mathcal{H})$, a Markov kernel $\mathbb{Q} : H \to \Delta(\mathcal{O})$ and a *prior* $\mu \in \Delta(\mathcal{H})$ such that

$\mathbb{P}^{\mathsf{X}_1\mathsf{X}_2\mathsf{X}_3} =$  (3.8)


Defining the hypothesis $\mathsf{H} : O \mapsto H$ such that $\mathbb{P}^{\mathsf{H}} = \mu$ and $\mathbb{P}^{\mathsf{O}|\mathsf{H}} = \mathbb{Q}$, $(\mathbb{Q}, \mathsf{H}, \mathsf{O})$ is a statistical model.

In the following section, we extend this result to the case of see-do forecasts. We first consider a model that is exchangeable in the observations only, and then introduce the notion of functional exchangeability which is a generalisation of exchangeability to Markov kernels. Finally, we prove a representation theorem for see-do forecasts that are both exchangeable in observations and functionally exchangeable in consequences.

**Definition 3.2.3** (Permutations and swaps)**.** A *finite permutation* $\rho'$ on $B \subseteq \mathbb{N}$ is a map $B \to B$ such that there is some finite $A \subset B$ for which $\rho'|_A : A \to A$ is an invertible function and $\rho'|_{B\setminus A} = \mathrm{Id}_{B\setminus A}$.

Given measureable space $(E, \mathcal{E})$ and a set of random variables $\{\mathsf{X}_i | i \in B\}$ the swap function $\rho^{\mathsf{X}} : E \to E$ associated with a finite permutation $\rho : B \to B$ and the random variables $\{\mathsf{X}_i\}_B$ is a $\mathcal{E}$ measurable function which has the property $\mathsf{X}_i \circ \rho^{\mathsf{X}} = \mathsf{X}_{\rho'(i)}$ for all $i \in B$, and for any $\mathsf{Y} : E \to Y$ with $\mathsf{Y}(\mathsf{X}^{-1}(A)) = Y$ for any $A \in \mathcal{E}$, $Y \circ \rho^{\mathsf{X}} = \mathsf{Y}$. This swap function also has an associated Markov kernel $R := \mathbb{F}_{\rho^{\mathsf{X}}}$.

For example, if $E = Y \times X_1^{|B|}$ and $\mathsf{X}_i : E \to X_1$ projects the $i$-th "x" element of the space $(y, x_1, ..., x_i, ...) \mapsto x_i$, then for some finite permuation $\rho$ the associated swap is the the fuction $\rho^{\mathsf{X}} : (y, x_1, ..., x_i, ...) \mapsto (y, x_{\rho'(1)}, ..., x_{\rho'(i)}, ...)$.

$$R \underline{\otimes}_{i\in B} \mathbb{F}_{\mathsf{X}_i} = \underline{\otimes}_{i\in B} R\mathbb{F}_{\mathsf{X}_i} \tag{3.9}$$
$$= \underline{\otimes}_{i\in B} \mathbb{F}_{\mathsf{X}_{\rho(i)}} \tag{3.10}$$

Where line 3.9 follows from the fact that deterministic kernels commute with the split map (2.167), and line 3.10 follows from the fact that for two functional kernels

$$(\mathbb{F}_f\mathbb{F}_g)_x(A) = \int_X (\mathbb{F}_g)_y(A) d(\mathbb{F}_f)_x(y) \tag{3.11}$$
$$= \int_X \delta_{g(y)}(A) d\delta_{f(x)}(y) \tag{3.12}$$
$$= \delta_{g(f(x))}(A) \tag{3.13}$$
$$= (\mathbb{F}_{g\circ f})_x(A) \tag{3.14}$$

lemmafy, move to chapter 2

**Definition 3.2.4** (Partial frequencies)**.** Given a standard measurable space $(E, \mathcal{E})$ along with random variables $\mathsf{X}_i : E \to X_1$ for each $i \in \mathbb{N}$, for $A \in \mathcal{X}_1$ and $m \leq n \in \mathbb{N}$ define the size $n$, $m$-tuple *partial frequency of* $A$ with respect to $\mathsf{X} := \underline{\otimes}_{i\in\mathbb{N}}\mathsf{X}_i$ to be $\mathsf{Z}_A^{m,n} := \frac{(n-m)!}{n!}\sum_{I\subset[n]}\prod_{i\in I}\mathbb{1}_A \circ \mathsf{X}_i$ where $I$ ranges over all $m$-sized ordered subsets of $n$.

Define the $m$-tuple *relative frequency of* $A$ with respect to $\mathsf{X}$ to be $\mathsf{Z}_A^{m,\infty} := \lim_{n\to\infty}\frac{(n-m)!}{n!}\sum_{I\in[n]}\prod_{i\in I}\mathbb{1}_A \circ \mathsf{X}_i$.

Given a countable set $\mathcal{G}$ generating $\mathcal{X}_1$ (i.e. $\sigma(\mathcal{G}) = \mathcal{X}_1$), define the $m$-tuple *relative frequency of* $\mathsf{X}$ to be $\mathsf{Z}^{m,\infty} := \underline{\otimes}_{A\in G}\mathsf{Z}_A^{m,\infty}$, if such a limit exists for all $A$.

For the special case of $m = 1$, let $\mathsf{Z} := \mathsf{Z}^{1,\infty}$

**Definition 3.2.5** (Exchangeable $\sigma$-algebra)**.** Given a set of random variables $\mathsf{X}_i : E \to X_1$ for each $i \in \mathbb{N}$, with $X = X_1^{\mathbb{N}}$, a *n-place symmetric function* $f : X \to W$ is a function for which $f = f \circ \rho$ for any permuation $\rho : \mathbb{N} \to \mathbb{N}$ such that $i > n \implies \rho(i) = i$.

The *n-place exchangeable $\sigma$-algebra* (with respect to the random variable $\mathsf{X}_i$), $\mathcal{H}^n$, is the $\sigma$-algebra generated by all n-place symmetric functions, and $\sigma H = \cap_{n\in\mathbb{N}}\mathcal{H}^n$.

For standard measurable $(E, \mathcal{E})$ and $n \in \mathbb{N}$, a size $n$ swap function $\rho^{\mathsf{X}n} : E \to E$ is a swap function associated with a permutation $\rho'^n$ with the property $i > n \implies \rho'^n(i) = i$. An $n$-symmetric set $S \subset E$ has the property $\rho^{\mathsf{X}n}(S) = S$ for all size $n$ swap functions $\rho^{\mathsf{X}n}$. Define the symmetric sets $\mathcal{S}^n$ $n$-symmetric sets, and $\mathcal{S} = \cap_{n\in\mathbb{N}}\mathcal{S}^n$. Given random variables $\mathsf{X}_i : E \to X_1$ for each $i \in \mathbb{N}$, the size $n$ *exchangeable $\sigma$-algebra* with respect to $\mathsf{X} := \underline{\otimes}_{i\in\mathbb{N}}\mathsf{X}_i$, denoted $\mathcal{H}^n \subset \sigma(\mathsf{X})$, is the $\sigma$-algebra generated by $\{\mathsf{X}^{-1}(A)|A \in \mathcal{X}\} \cap S^n$.

**Lemma 3.2.6.** *The size n exchangeable sigma algebra $\mathcal{H}^n$ on $(E, \mathcal{E})$ with respect to $\mathsf{X} := \underline{\otimes}_{i\in\mathbb{N}}\mathsf{X}_i$ has the property $\rho^{\mathsf{X}n}(A) = A$ for all $A \in \mathcal{H}^n$, and all size n swap functions $\rho^{\mathsf{X}n}$.*

*Proof.* Let $W_f$ be the codomain of a function $f$, and $\mathcal{W}_f$ its $\sigma$-algebra. $\mathcal{H}^n$ is generated by $\mathcal{G}^n = \{f^{-1}(A)|A \in \mathcal{W}_f, f \text{ n-place symmetric}\}$. By the definition of n-place symmetric functions, any set of the form $f^{-1}(A) = (f \circ \rho^{\mathsf{X}})^{-1}(A) = (\rho^{\mathsf{X}})-1(f^{-1}(A))$. Because every $n$-place permutation $\rho$ has an inverse $\rho^{-1}$ that is also an $n$-place permutation, all sets in $B \in \mathcal{G}$ have the property $\rho^{\mathsf{X}}(B) = B$ for all $n$-place permuations $\rho$.

Define $\mathcal{S}^n$ to be all subsets of $E$ such that for $B \in \mathcal{S}^n$, $n$-place permuations $\rho$, $\rho^{\mathsf{X}}(B) = B$. $\mathcal{S}^n$ is a $\sigma$-algebra, and it contains $\mathcal{G}^n$, so it also contains $\mathcal{H}^n$.

Take $A \in \mathcal{S}^n$. By assumption, for any $\omega \in A$, $\rho^{\mathsf{X}}(\omega) \in A$ for all size $n$ swap functions $\rho^{\mathsf{X}}$. Consider $\omega \notin A$, and suppose there is some swap function $\rho^{\mathsf{X}}$ such that $\rho^{\mathsf{X}}(\omega) \in A$. By definition, the permutation $\rho$ has an inverse $\rho^{-1}$ which is also a size $n$ permutation. By construction, $(\rho^{-1})^{\mathsf{X}}$ is also the inverse of $\rho^{\mathsf{X}}$. Thus $(\rho^{-1})^{\mathsf{X}}(\omega) \notin A$ and so $\omega \notin A$, a contradiction. Thus $E \setminus A \in \mathcal{G}^n$.

For any invertible function $f : E \to E$, $f(E) = E$. Thus $E \in \mathcal{G}^n$.

Finally, for a countable collection $A_1, A_2, ...$ and any size $n$ swap function $\rho^{\mathsf{X}}$, $\rho^{\mathsf{X}}(\cup_{i=1}^{\infty} A_i) = \cup_{i=1}^{\infty} \rho^{\mathsf{X}}(A_i) = \cup_{i=1}^{\infty} A_i$. Thus $\mathcal{S}^n$ is a $\sigma$-algebra, and by the monotone class theorem it contains $\mathsf{H}^n$. □

**Definition 3.2.7** (Exchangeability)**.** Given a see-do forecast $(\mathbb{T}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ with the property that $\mathsf{X} := \underline{\otimes}_{i \in A} \mathsf{X}_i$ for some set of random variables $\{\mathsf{X}_i | i \in A\}$ all taking values in $X_1$ where $A \subseteq \mathbb{N}$.

If for every finite $B \subset A$ and every permutation $\rho' : B \to B$ of $B$ we have $\mathbb{T}\rho = \mathbb{T}$, where $\rho$ is the swap associated with $\rho'$ and $\{\mathsf{X}_i | i \in B\}$, then $(\mathbb{T}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ is *exchangeable* with respect to $\{\mathsf{X}_i | i \in A\}$.

If $A$ is an infinite set then $\mathbb{T}$ is *infinitely exchangeable*, and if $\mathbb{T} = \mathbb{S}(\mathrm{Id}_X \otimes * \otimes \mathrm{Id}_Y)$ for some infinitely exchangeable $(\mathbb{S}, \mathsf{D}, \mathsf{X}', \mathsf{Y}')$, then $\mathbb{T}$ is infinitely exchangeably extendable.

Note that $\mathbb{T}R^{\mathsf{X}_i|\mathsf{D}} = \mathbb{T}^{\mathsf{X}_{\rho(i)}|\mathsf{D}}$.

This implies the usual notion of exchangeability if we take $Y = \{*\}$ (that is, if we assume the consequences are trivial), as by assumption $\mathsf{X}$ is independent of $\mathsf{D}$.

**Lemma 3.2.8** (Infinitely exchangeably extendable forecasts)**.** *Given a forecast $(\mathbb{P}, \mathsf{X})$ where $\mathsf{X} = \underline{\otimes}_{i \in A} \mathsf{X}_i$ for some $\{\mathsf{X}_i | i \in \mathbb{N}\}$ and $X$ is standard measurable, if $\mathbb{P}$ is exchangeable with respect to $\mathsf{X}$ then there exists a function $f : X \to H$ such that, defining $\mathsf{Z} : f \circ \mathsf{X}$:*

- $\mathsf{X}_i \perp\!\!\!\perp \mathsf{X}_{A \setminus \{i\}} | \mathsf{Z}$ *for all* $i \in A$
- $\mathbb{P}^{\mathsf{X}_i|\mathsf{Z}} = \mathbb{P}^{\mathsf{X}_j|\mathsf{Z}}$ *for all* $i, j \in A$

*Call* $\mathsf{Z}$ *the* hypothesis.

*Proof.* Without loss of generality, assume $X_1 = [0, 1]$, $\mathcal{X} = \mathcal{B}([0, 1])$ and $\mathsf{X} = [0, 1]^{\mathbb{N}}$.

Let $\mathbb{Q}$ be the rationals between $[0, 1]$ and define $\mathsf{Z}_q^n : D \times X \times Y \to [0, 1]$ by $\omega \mapsto \frac{1}{n} \sum_i^n \mathbb{1}_{[0,q)}(\mathsf{X}_i(\omega))$ Let $\mathcal{Z}_q^n = \sigma(\mathsf{Z}_q^n)$, i.e. $\mathsf{Z}^n$ is a 1-tuple partial frequency as in Definition 3.2.4.

$\mathsf{Z}^n \circ \rho^{\mathsf{X}n} = \mathsf{Z}^n$ for any size $n$ swap function $\rho^{\mathsf{X}n}$, so $\mathsf{Z}^n$ is $\mathcal{H}^n$-measurable.

Let $\rho'_{ij} : \mathbb{N} \to \mathbb{N}$ swaps indices $i$ and $j$ for some $i, j \in [n]$ and otherwise acts as the identity. $\rho_{ij} : D \times X \times Y \to \Delta(\mathcal{D} \times \mathcal{X} \times \mathcal{Y})$ is the swap kernel associated with $\rho'_{ij}$ and $\{\mathsf{X}_i | i \in \mathbb{N}\}$, and $\rho_{ij}^{\mathsf{X}}$ the function associated with $\rho_{ij}$. For any $m$, $n$, $A \in \mathcal{H}^n$, $d \in D$:

$$\int_A \mathsf{Z}_q^n(\omega)d\mathbb{P}(\omega) = \int_A \frac{1}{n}\sum_i^n (\mathbb{1}_{[0,q)} \circ \mathsf{X}_i)(\omega)d\mathbb{P}(\omega) \tag{3.15}$$

$$= \frac{1}{n}\sum_i^n \int_{(\rho_{ij}^{\mathsf{X}})^{-1}\rho_{ij}^{\mathsf{X}}(A)} (\mathbb{1}_{[0,q)} \circ \mathsf{X}_i)(\omega)d\mathbb{P}\rho_{ij}(\omega) \tag{3.16}$$

$$= \frac{1}{n}\sum_i^n \int_{\rho_{ij}^{\mathsf{X}}(A)} (\mathbb{1}_{[0,q)} \circ \mathsf{X}_i \circ \rho_{ij})(\omega)d\mathbb{P}(\omega) \tag{3.17}$$

$$= \frac{1}{n}\sum_i^n \int_A (\mathbb{1}_{[0,q)} \circ \mathsf{X}_j)(\omega)d\mathbb{P}(\omega) \tag{3.18}$$

$$= \int_A (\mathbb{1}_{[0,q)} \circ \mathsf{X}_j)(\omega)d\mathbb{P}(\omega) \tag{3.19}$$

Where line 3.16 follows from exchangeability of $\mathbb{T}$ and invertibility of $\rho_{ij}$. Line 3.17 follows from the fact that $\mathbb{P}\rho_{ij}$ is the pushforward measure of $\mathbb{P}$ with respect to $\rho_{ij}^{\mathsf{X}}$ and 3.18 uses the fact that $\rho(A) = A$ for all $A \in \mathsf{Z}_q^n$ and all permutations $\rho$.

From Equation 3.19, we have for all $n$, $A \in \mathcal{H}^{n+1}$

$$\int_A \mathsf{Z}_q^{n+1}(\omega)d\mathbb{P}(\omega) = \int_A \mathsf{Z}_q^n(\omega)d\mathbb{P}(\omega) \tag{3.20}$$

Because $\mathsf{Z}_q^{n+1}$ is $\mathcal{H}^{n+1}$ measurable, $\mathsf{Z}_q^{n+1} = \mathbb{E}[\mathsf{Z}_q^n|\mathcal{H}^{n+1}]$.

Thus the sequence $[\mathsf{Z}_q^1, \mathsf{Z}_q^2, ...]$ is a backwards martingale with respect to the reversed filtration $\mathcal{H}^1 \supset \mathcal{H}^2 \supset ... \supset \mathcal{H}^3$.

Furthermore, for all $n \in \mathbb{N}$, $\sup_\omega |\mathsf{Z}^n(\omega)| \leq 1$ so the sequence is also uniformly integrable. Thus it goes almost surely to the limit $\mathsf{Z}_q$, and for all $A \in \mathcal{H}$

$$\lim_{n\to\infty} \int_A \mathsf{Z}_q^n(\omega)d\mathbb{P}(\omega) = \int_A \mathsf{Z}_q(\omega)d\mathbb{P}(\omega) \tag{3.21}$$

Finally, because for all $n \in \mathbb{N}$, all $j \in [n]$ and all $A \in \mathcal{H}^n$ we also have

$$\int_A \mathbb{1}_{[0,q)}(\mathsf{X}_j(\omega))d\mathbb{P}(\omega) = \int_A \mathsf{Z}_q^n(\omega)d\mathbb{P}(\omega) \tag{3.22}$$

it follows that for all $A \in \mathcal{H}$, $j \in \mathbb{N}$

$$\int_A \mathsf{Z}_q(\omega)d\mathbb{P}(\omega) = \int_A \mathbb{1}_{[0,q)}(\mathsf{X}_j(\omega))d\mathbb{P}(\omega) \tag{3.23}$$

[Çinlar (2011) Thm 4.7.]. Thus $\mathsf{Z}_q = \mathbb{E}[\mathbb{1}_{[0,q)} \circ \mathsf{X}_j|\mathcal{H}]$ for all $j \in \mathbb{N}$. This implies $\mathsf{Z}_q$ is a version of $\mathbb{E}[\mathbb{1}_{[0,q)} \circ \mathsf{X}_j|\mathcal{H}]$.

Define $\mathsf{Z} = \underline{\otimes}_q \in \mathbb{Q}\mathsf{Z}_q$. As $\sigma(\mathsf{Z}) \subset \mathcal{H}$, Equation 3.23 establishes in addition that $\mathsf{Z}_q$ is a version of $\mathbb{E}[\mathbb{1}_{[0,q)} \circ \mathsf{X}_j | \sigma(Z)]$.

By the definition of conditional expectation, for any version of $\mathbb{P}^{\mathsf{X}_j|\mathsf{ZD}}_{\mathsf{Z}(\omega)}([0,q))$ we have

$$\mathbb{P}^{\mathsf{X}_j|\mathsf{Z}}_{\mathsf{Z}(\omega)}([0,q)) = \mathbb{E}[\mathbb{1}_{[0,q)} \circ \mathsf{X}_j | \sigma(\mathsf{Z})](\omega) \tag{3.24}$$

$$\tag{3.25}$$

$\mathbb{P}$-almost surely.

Furthermore, the measure $\mathbb{P}^{\mathsf{X}_j|\mathsf{Z}}_h$ is uniquely defined by its value on $[0,q)$ for all $q \in \mathbb{Q}$. Thus for all $i, j \in \mathbb{N}$ we have

$$\mathbb{P}^{\mathsf{X}_j|\mathsf{H}} = \mathbb{P}^{\mathsf{X}_i|\mathsf{H}} \tag{3.26}$$

Completing the proof of property 1.

Next, we will show $\mathsf{X}_i \perp\!\!\!\perp_{\mathbb{T}} \mathsf{X}_{\mathbb{N}\setminus\{i\}} | \mathsf{H}$.

Let $\mathsf{Z}^{m,n}_q$ be a partial frequency as in Definition 3.2.4 where $q$ stands for the set $[0,q)$. $\mathsf{Z}^{m,n}_q \circ \rho^{\mathsf{X}n} = \mathsf{Z}^{m,n}_q$ for all size $n$ swap functions so $\mathsf{Z}^{m,n}_q$ is $\mathcal{H}^n$ measurable.

Let $J \subset [n]$ be some set of $m$ elements from $n$ and $\rho'_{IJ}$ be a permutation that sends the elements of $I \subset [n]$ to $J$.

$$\int_A \mathsf{Z}^{m,n}_q(\omega) d\mathbb{P}(\omega) = \int_A \frac{(n-m)!}{n!} \sum_{I\subset[n]} \prod_{i\in I} (\mathbb{1}_{[0,q)} \circ \mathsf{X}_i)(\omega) d\mathbb{P}(\omega) \tag{3.27}$$

$$= \frac{(n-m)!}{n!} \sum_{I\subset[n]} \int_{(\rho^{\mathsf{X}}_{IJ})^{-1}\rho^{\mathsf{X}}_{IJ}(A)} \prod_{i\in I} (\mathbb{1}_{[0,q)} \circ \mathsf{X}_i)(\omega) d\mathbb{P}\rho_{IJ}(\omega) \tag{3.28}$$

$$= \frac{(n-m)!}{n!} \sum_{I\subset[n]} \int_{\rho^{\mathsf{X}}_{IJ}(A)} \prod_{i\in I} (\mathbb{1}_{[0,q)} \circ \mathsf{X}_i \circ \rho_{IJ})(\omega) d\mathbb{P}(\omega) \tag{3.29}$$

$$= \frac{(n-m)!}{n!} \sum_{I\subset[n]} \int_A \prod_{j\in J} (\mathbb{1}_{[0,q)} \circ \mathsf{X}_j)(\omega) d\mathbb{P}(\omega) \tag{3.30}$$

$$= \int_A \prod_{j\in J} (\mathbb{1}_{[0,q)} \circ \mathsf{X}_j)(\omega) d\mathbb{P}(\omega) \tag{3.31}$$

From Equation 3.31, we have for all $n$, $m < n$, $A \in \mathcal{H}^{n+1}$

$$\int_A \mathsf{Z}^{m,n+1}_q(\omega) d\mathbb{P}(\omega) = \int_A \mathsf{Z}^{m,n}_q(\omega) d\mathbb{P}(\omega) \tag{3.32}$$

Because $\mathsf{Z}_q^{m,n+1}$ is $\mathcal{H}^{n+1}$ measurable, $\mathsf{Z}_q^{m,n+1} = \mathbb{E}[\mathsf{Z}_q^{m,n}|\mathcal{H}^{n+1}]$.

Thus the sequence $[\mathsf{Z}_q^{m,1}, \mathsf{Z}_q^{m,2}, ...]$ is a backwards martingale with respect to the reversed filtration $\mathcal{H}^1 \supset \mathcal{H}^2 \supset ... \supset \mathcal{H}^3$.

Furthermore, for all $n \in \mathbb{N}$, $\sup_\omega |\mathsf{Z}_q^{m,n}(\omega)| \leq 1$ so the sequence is also uniformly integrable. Thus it goes almost surely to a limit $\mathsf{Z}_q^{m,\infty}$, and for all $A \in \mathcal{H}$

$$\lim_{n\to\infty} \int_A \mathsf{Z}_q^{m,n}(\omega) d\mathbb{P}(\omega) = \int_A \mathsf{Z}_q^{m,\infty}(\omega) d\mathbb{P}(\omega) \tag{3.33}$$

$$\implies \mathsf{Z}_q^{m,n} = \mathbb{E}[\prod_{j\in J} \mathbb{1}_{[0,q)} \circ \mathsf{X}_j | \mathcal{H}] \tag{3.34}$$

Let $[n]^m_{\text{rep}}$ be the set of all $m$ length sequences of elements of $[n]$ with repeats. Note that $\lim_{n\to\infty} \frac{(n-m)!}{n!} |[n]^m_{\text{rep}}| = \lim_{n\to\infty} \frac{n^m (n-m)!}{n!} - 1 = 0$

$$\mathbb{E}[\prod_{j\in J} \mathbb{1}_{[0,q)} \circ \mathsf{X}_j | \mathcal{H}](\omega) = \lim_{n\to\infty} \frac{(n-m)!}{n!} \sum_{I\subset[n]} \prod_{i\in I} (\mathbb{1}_{[0,q)} \circ \mathsf{X}_i) \tag{3.35}$$

$$= \lim_{n\to\infty} \frac{1}{n^m} \sum_{I\subset[n]} \prod_{i\in I} (\mathbb{1}_{[0,q)} \circ \mathsf{X}_i) \tag{3.36}$$

$$= \lim_{n\to\infty} \frac{1}{n^m} \left[ \sum_{i_1\in[n]} ... \sum_{i_m\in[n]} (1_{[0,q)} \circ \mathsf{X}_{i_k}) - \sum_{J\subset[n]^m_{\text{rep}}} \prod_{i\in I} (\mathbb{1}_{[0,q)} \circ \mathsf{X}_i) \right] \tag{3.37}$$

$$= \prod_{k\in[m]} \lim_{n\to\infty} \frac{1}{n} \sum_{i_k\in[n]} (1_{[0,q)} \circ \mathsf{X}_{ik}) \tag{3.38}$$

$$= \prod_{k\in[m]} \mathbb{E}[\mathbb{1}_{[0,q)} \circ \mathsf{X}_j | \mathcal{H}] \tag{3.39}$$

$$= \prod_{k\in J} \mathbb{E}[\mathbb{1}_{[0,q)} \circ \mathsf{X}_j | \mathcal{H}] \tag{3.40}$$

$$= \prod_{k\in J} \mathsf{Z}_q \tag{3.41}$$

Because $\mathbb{E}[\prod_{j\in J} \mathbb{1}_{[0,q)} \circ \mathsf{X}_j | \mathcal{H}]$ is $\mathsf{Z}$-measurable we also have

$$\mathbb{E}[\prod_{j\in J} \mathbb{1}_{[0,q)} \circ \mathsf{X}_j | \sigma(\mathsf{Z})] = \prod_{k\in J} \mathsf{Z}_q \tag{3.42}$$

By the definition of conditional expectation, for all $J \subset \mathbb{N}$

$$\mathbb{P}^{\mathsf{X}_J|\mathsf{H}}_{\mathsf{H}(\omega)}(\times_{j\in J}[0,q)) = \mathbb{E}[\prod_{j\in J} \mathbb{1}_{[0,q)} \circ \mathsf{X}_j|\mathcal{H}](\omega) \tag{3.43}$$

$$= \prod_{k\in J} \mathbb{P}^{\mathsf{X}_j|\mathsf{H}}_{\mathsf{H}(\omega)}([0,q)) \tag{3.44}$$

And thus $\mathsf{X}_i \perp\!\!\!\perp \mathsf{X}_{\mathbb{N}\setminus\{i\}}|\mathsf{Z}$ for all $i \in \mathbb{N}$. □

**Lemma 3.2.9** (Independent and identically distributed random variables)**.** *Suppose we have a forecast* $(\mathbb{P}, \mathsf{X})$ *where* $\mathsf{X} = \underline{\otimes}_{i\in A}\mathsf{X}_i$ *for some* $\{\mathsf{X}_i|i \in \mathbb{N}\}$ *and* $X$ *is standard measurable, and* $\mathbb{P}$ *is exchangeable with respect to* $\mathsf{X}$*. Furthermore, suppose we have* $\mathsf{V} = \underline{\otimes}_{i\in\mathbb{N}} f \circ \mathsf{X}_i$ *for some measurable* $f : X_1 \to V_1$ *such that* $\mathsf{V}$ *is independent and identically distributes - that is,* $\mathbb{P}^{\mathsf{V}} = \otimes_{i\in\mathbb{N}}\mathbb{P}^{\mathsf{V}_1}$*. Then, letting* $\mathsf{Z}$ *be the hypothesis from Lemma 3.2.8,* $\mathsf{V} \perp\!\!\!\perp \mathsf{Z}$*.*

*Proof.* We have by Lemma 3.2.8 that fo rany measurable $g : Z_1 \to \mathbb{R}$,

$$\mathbb{E}[g(\mathsf{V}_1)|\sigma(\mathsf{Z})] = \lim_{n\to\infty} \frac{1}{n}\sum_{i}^{n} g(\mathsf{Z}_i) \tag{3.45}$$

However, by the strong law of large numbers (Çinlar (2011), pg 119), because the $\mathsf{V}_i$ are independent and identically distributed

$$\mathbb{E}[g(\mathsf{V}_1)] \overset{a.s.}{=} \lim_{n\to\infty} \frac{1}{n}\sum_{i}^{n} g(\mathsf{Z}_i) \tag{3.46}$$

$$= \mathbb{E}[g(\mathsf{V}_1)|\sigma(\mathsf{Z})] \tag{3.47}$$

Thus $\mathsf{V} \perp\!\!\!\perp \mathsf{Z}$.

□

**Lemma 3.2.10** (Representation of infinitely exchangeably extendable see-do forecasts)**.** *Given a see-do forecast* $(\mathbb{T}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ *where* $\mathsf{X} = \underline{\otimes}_{i\in\mathbb{N}}\mathsf{X}_i$ *for some* $\{\mathsf{X}_i|i \in \mathbb{N}\}$ *and* $X \times Y$ *is standard measurable, if* $(\mathbb{T}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ *is infinitely exchangeable with respect to* $\{\mathsf{X}_i|i \in \mathbb{N}\}$ *then there exists a function* $f : X \to Z$ *such that, defining* $\mathsf{Z} : f \circ \mathsf{X}$*:*

- $\mathsf{X}_i \perp\!\!\!\perp \mathsf{X}_{\mathbb{N}\setminus\{i\}}|\mathsf{Z}$ *for all* $i \in A$
- $\mathbb{T}^{\mathsf{X}_i|\mathsf{Z}} = \mathbb{T}^{\mathsf{X}_j|\mathsf{Z}}$ *for all* $i, j \in A$
- $\mathsf{Y} \perp\!\!\!\perp \mathsf{X}|\mathsf{Z}\underline{\otimes}\mathsf{D}$

*Proof.* Without loss of generality, assume $X_1 = Y = [0, 1]$.

$\mathbb{T}$ is a see-do forecast, so $\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}$. Thus there exists a marginal $\mathbb{T}^{\mathsf{X}}$ independent of $\mathsf{D}$.

From Lemma 3.2.8, $\mathbb{T}^{\mathsf{X}_i|\mathsf{Z}} = \mathbb{T}^{\mathsf{X}_j|\mathsf{Z}}$ for all $i, j \in \mathbb{N}$ and $\mathsf{X}_i \perp\!\!\!\perp \mathsf{X}_{\mathbb{N}\setminus\{i\}}|\mathsf{Z}$.

We will show $\mathsf{Y} \perp\!\!\!\perp \mathsf{X}|\mathsf{Z}\underline{\otimes}\mathsf{D}$.

For any swap function $\rho^{\mathsf{X}}$ there is, by definition, a permutation of indices $\rho'$ such that $\mathsf{X} \circ \rho^{\mathsf{X}}(\omega) = [\mathsf{X}_{\rho'(1)}(\omega), \mathsf{X}_{\rho'(2)}(\omega), ...]$. Define $\rho'' : [0,1]^{\mathbb{N}} \to [0,1]^{\mathbb{N}}$ to be the bijective map that performs the permutation in the codomain of $\mathsf{X}$, i.e. $\mathsf{X} \circ \rho^{\mathsf{X}} = \rho'' \circ \mathsf{X}$.

Consider some $B \in \mathcal{B}([0,1])^{\mathbb{N}}$ and its preimage $\mathsf{X}^{-1}(B) = \{\omega|\mathsf{X}(\omega) \in B\}$, and some finite swap function $\rho^{\mathsf{X}}$. Then there exists $\rho''(B) \in [0,1]^{\mathbb{N}}$ such that $(\mathsf{X}\circ\rho^{\mathsf{X}})^{-1}(\rho''(B)) = \{\omega|\mathsf{X}(\rho^{\mathsf{X}}(\omega)) \in \rho''(B)\} = \{\omega|\mathsf{X} \in B\}$. Thus $\sigma(\mathsf{X}) = \sigma(\mathsf{X}\circ\rho^{\mathsf{X}})$ for any finite swap function $\rho^{\mathsf{X}}$.

Because $E = X \times Y$ and $\mathsf{X}$ forgets the $Y$-component of any $(x, y) \in E$, for any $A, B \in \mathcal{B}([0,1])$ there is some $C \subset X$ such that $\mathbb{1}_A \circ \mathsf{Y}(\mathsf{X}^{-1}(B)) = \mathbb{1}_A \circ \mathsf{Y}(C \times Y) = [0,1]$. Thus for any finite swap function $\rho^{\mathsf{X}}$, $\mathbb{1}_A \circ \mathsf{Y} \circ \rho^{\mathsf{X}} = \mathbb{1}_A \circ \mathsf{Y}$.

For $A \in \sigma(\mathsf{X})$:

$$\int_A \mathbb{E}[\mathbb{1}_A \circ \mathsf{Y}|\sigma(\mathsf{X})] \circ \rho^{\mathsf{X}} d\mathbb{T}_d = \int_{\rho^{\mathsf{X}}(A)} \mathbb{E}[\mathbb{1}_A \circ \mathsf{Y}|\sigma(\mathsf{X})] d(\mathbb{T}R^{-1})_d \tag{3.48}$$

$$= \int_{\rho^{\mathsf{X}}(A)} \mathbb{E}[\mathbb{1}_A \circ \mathsf{Y}|\sigma(\mathsf{X})] d\mathbb{T}_d \tag{3.49}$$

$$= \int_{\rho^{\mathsf{X}}(A)} \mathbb{1}_A \circ \mathsf{Y} d\mathbb{T}_d \tag{3.50}$$

$$= \int_A \mathbb{1}_A \circ \mathsf{Y} d\mathbb{T}_d \tag{3.51}$$

It follows that $\mathbb{E}[\mathbb{1}_A \circ \mathsf{Y}|\sigma(\mathsf{X})] \circ \rho^{\mathsf{X}}$ is a version of $\mathbb{E}[\mathbb{1}_A \circ \mathsf{Y}|\sigma(\mathsf{X})]$. Because there are only a countable number of finite swap functions, it also follows that a version of $\mathbb{E}[\mathbb{1}_A \circ \mathsf{Y}|\sigma(\mathsf{X})]$ that is $\mathcal{H}$-measurable exists (i.e. for which $\mathbb{E}[\mathbb{1}_A \circ \mathsf{Y}|\sigma(\mathsf{X})] \circ \rho^{\mathsf{X}} = \mathbb{E}[\mathbb{1}_A \circ \mathsf{Y}|\sigma(\mathsf{X})]$, for all $\rho^{\mathsf{X}}$).

Consider also for some swap function $\rho^{\mathsf{X}}$, $B \in \mathcal{B}([0,1])$:

$$\int_{\mathsf{X}_i^{-1}(B)} \mathbb{E}[\mathsf{V}|\sigma(\mathsf{X}_j)] \circ \rho^{\mathsf{X}}_{ji} d\mathbb{T}_d = \int_{\rho^{\mathsf{X}}_{ji}(\mathsf{X}_i^{-1}(B))} \mathbb{E}[\mathsf{V}|\sigma(\mathsf{X}_j)] d(\mathbb{T}R_{ij})_d \tag{3.52}$$

$$= \int_{\mathsf{X}_j^{-1}(B)} \mathbb{E}[\mathsf{V}|\sigma(\mathsf{X}_j)] d\mathbb{T}_d \tag{3.53}$$

$$= \int_{\mathsf{X}_j^{-1}(B)} \mathsf{V} d\mathbb{T}_d \tag{3.54}$$

$$= \int_{\rho^*_{ji}(\mathsf{X}_j^{-1}(B))} \mathsf{V} \circ \rho^{\mathsf{X}}_{ji} d(\mathbb{T}R_{ji})_d \tag{3.55}$$

$$= \int_{\mathsf{X}_i^{-1}(B)} \mathsf{V} d\mathbb{T}_d \tag{3.56}$$

Thus $\mathbb{E}[\mathsf{V}|\sigma(\mathsf{X}_j)] \circ \rho^{\mathsf{X}}_{ji}$ is a version of $\mathbb{E}[\mathsf{V}|\sigma(\mathsf{X}_i)]$.

Define $\mathsf{W}_q^n := \frac{1}{n}\sum_{i\in[n]} \mathbb{E}[\mathbb{1}_{[0,q]}\mathsf{Y}|\sigma(\mathsf{X}_i)]$. Note that for any size $n$ swap function $\rho^{\mathsf{X}}$, $\mathsf{W}_q^n \circ \rho^{\mathsf{X}} = \mathsf{W}_q^n$, therefore $\mathsf{W}_q^n$ is $\mathcal{H}^n$-measurable.

Consider $\omega, \omega'$ such that $\mathsf{W}_q^n(\omega) \neq \mathsf{W}_q^n(\omega')$. Then there exists no size $n$ swap function $\rho^{\mathsf{X}}$ such that $\mathsf{X}_{[n]}(\omega) = \mathsf{X}_{[n]}(\rho^{\mathsf{X}}(\omega'))$. Without loss of generality, suppose $\mathsf{X}_1(\omega) \leq \mathsf{X}_2(\omega) \leq ... \leq \mathsf{X}_n(\omega)$ and $\mathsf{X}_1(\omega') \leq \mathsf{X}_2(\omega') \leq ... \leq \mathsf{X}_n(\omega')$. Then there is some first index $j$ for which $\mathsf{X}_j(\omega) > \mathsf{X}_j(\omega')$, and some and some rational $r$ such that $\mathsf{X}_j(\omega) > r > \mathsf{X}_j(\omega')$. Then $\sum_i^n \mathbb{1}_{[0,q)}(\mathsf{X}_i(\omega)) > \sum_i^n \mathbb{1}_{[0,q)}(\mathsf{X}_i(\omega'))$ and so $\mathsf{Z}^n(\omega) \neq \mathsf{Z}^n(\omega')$ also.

Thus $\mathsf{W}_q^n$ is $\sigma(\mathsf{Z}^n)$ measurable. Define $\mathcal{I}^n := \vee_{n\to\infty}\sigma(\mathsf{Z}^n)$. Then $\mathcal{I}^1 \supset \mathcal{I}^2 \supset ... \supset \sigma(\mathsf{Z}) = \cap_i^\infty \mathcal{I}^i$. In addition, but the $\mathcal{H}^n$-measurability of $\mathsf{Z}^{>n}$, $\mathcal{I}^n \subset \mathcal{H}^n$ and $\mathcal{I} \subset \mathcal{H}$.

For $A \in \mathcal{I}^n$, $d \in D$:

$$\int_A \mathsf{W}_q^n d\mathbb{T}_d = \frac{1}{n}\sum_{i\in[n]} \int_A \mathbb{E}[\mathbb{1}_{[0,q)} \circ \mathsf{Y}|\sigma(\mathsf{X}_i)]d\mathbb{T}_d \tag{3.57}$$

$$= \frac{1}{n}\sum_{i\in[n]} \int_{\rho_{ij}^*(A)} \mathbb{E}[\mathbb{1}_{[0,q)} \circ \mathsf{Y}|\sigma(\mathsf{X}_i)] \circ \rho_{ij}^* d(\mathbb{T}R_{ij})_d \tag{3.58}$$

$$= \int_A \mathbb{E}[\mathbb{1}_{[0,q)} \circ \mathsf{Y}\sigma(\mathsf{X}_j)]d\mathbb{T}_d \tag{3.59}$$

$$= \int_A \mathbb{1}_{[0,q)} \circ \mathsf{Y} d\mathbb{T}_d \tag{3.60}$$

Where the last line follows from the fact that $\mathcal{I}^n \subset \sigma(\mathsf{X}_j)$. Therefore $\mathsf{W}_q^n$ is a version of $\mathbb{E}[\mathbb{1}_{[0,q)} \circ |\mathcal{I}^n]$. Furthermore $\mathsf{W}_q^1, ..., \mathsf{W}_q^n, ..$ is a backwards martingale with respect to $\mathcal{I}^1, ...., sigalgI^n, ..\mathcal{I}$ and so it goes to a limit $\mathsf{W}_q = \mathbb{E}[\mathbb{1}_{[0,q)} \circ \mathsf{V}|\sigma(\mathsf{Z})]$.

So $\mathsf{W}_q$ is a $\sigma(\mathsf{Z})$-measurable version of $\mathbb{E}[\mathbb{1}_{[0,q]} \circ \mathsf{Y}|\mathcal{H}]$ which is itself a version of $\mathbb{E}[\mathbb{1}_{[0,q]} \circ \mathsf{Y}|\sigma(\mathsf{X})]$. As before, for any $d \in D$, $\omega \in E$ we have

$$\mathbb{T}^{\mathsf{Y}|\mathsf{XZD}}_{\mathsf{D}(\omega),\mathsf{X}(\omega),\mathsf{Z}(\omega)}([0,q)) = \mathbb{E}[\mathbb{1}_{[0,q]} \circ \mathsf{Y}|\mathcal{H}](\omega) \tag{3.61}$$

$$= \mathbb{T}^{\mathsf{Y}|\mathsf{ZD}}_{\mathsf{D}(\omega),\mathsf{Z}(\omega)}([0,q)) \tag{3.62}$$

This shows that $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{X}|\mathsf{DZ}$, as desired.

□

Using Lemma 3.2.10 it is possible to prove a version of De Finetti's representation theorem, which is a well known result (de Finetti, 1992; Hewitt and Savage, 1955). We are interested in establishing an analogous result for consequence maps. This requires the assumption of *functional exchangeability.*

Functional exchangeability is a generalisation of exchangeablility to Markov kernels. It captures the intuition that if we swap the order of the outputs (say, $\mathsf{Y}_1, \mathsf{Y}_2$ is swapped to $\mathsf{Y}_2, \mathsf{Y}_1$), we need to make analagous exchange of choices

($\mathsf{D}_1$, $\mathsf{D}_2$ becomes $\mathsf{D}_2$, $\mathsf{D}_1$) in order to maintain the correspondence of choices and outputs.

**Definition 3.2.11** (Functional Exchangeability)**.**

Maybe include a diagram here? I think the pictorial representation helps with intuition, though it's hard to state as rigorously with pictures

Given a see-do forecast $(\mathbb{T}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ where $\mathsf{D} = \underline{\otimes}_{i\in A}\mathsf{D}_i$ and $\mathsf{Y} = \underline{\otimes}_{i\in A}\mathsf{X}_i$ for some random variables $\{\mathsf{D}_i\}_A, \{\mathsf{Y}_i\}_A$, for any perumtation $\rho : A \to A$ define the observation and choice swap function $\rho^{\mathsf{DY}}$ and the choice swap function $\rho^{\mathsf{D}}$ as in 3.2.3. Then $\mathbb{T}$ is functionally exchangeable with respect to $\mathsf{D}$ and $\mathsf{X}$ if $\mathbb{T} = R^{\mathsf{D}}\mathbb{T}R^{\mathsf{DY}}$.

$\mathbb{T}$ is infinitely functionally exchangeably extendable if here exists a do forecast $(\mathbb{T}', \mathsf{D}', \mathsf{X}')$ non-interfering and functionally exchangeable with respect to $\mathsf{D}' = \underline{\otimes}_{i\in\mathbb{N}}\mathsf{D}'_i$ and $\mathsf{Y}' = \underline{\otimes}_{i\in\mathbb{N}}\mathsf{Y}'_i$ such that for all $B \subset A$

$$\mathbb{T}'^{\mathsf{Y}'_B|\mathsf{D}'_B} = \mathbb{T}^{\mathsf{Y}_B|\mathsf{D}_B} \tag{3.63}$$

A see-do forecast that is infinitely exchangeably extendable with respect to $\mathsf{X}$ and infinitely functionally exchangeably extendable with respect to $\mathsf{D}$, $\mathsf{Y}$ is *doubly exchangeable* (we omit "infinitely extendable" for the sake of brevity).

The following lemma interprets an exchangeable probability measure as an exchangeable see-do forecast with observations and consequences interchanged. This is so we can re-use Lemma 3.2.10 without separately proving an almost identical result for probability measures.

**Lemma 3.2.12** (Functionally exchangeable see-do models with exchangeable choices induce exchangeable see do models)**.** *Given a see-do forecast* $(\mathbb{T}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ *functionally exchangeable with respect to* $\mathsf{Y} = \underline{\otimes}_{i\in A}\mathsf{Y}_i$ *and* $\mathsf{D} = \underline{\otimes}_{i\in A}\mathsf{D}_i$ *and some* $\mathbb{P}^{\mathsf{D}}$ *exchangeable with respect to* $\mathsf{D}$*, then* $\mathbb{P}\mathbb{T} \in \Delta(\mathcal{D} \otimes \mathcal{X} \otimes \mathcal{Y})$ *is exchangeable with respect to* $\mathsf{G} := \underline{\otimes}_i \in A\mathsf{Y}_i\underline{\otimes}\mathsf{D}_i$.

*Furthermore, defining the trivial choice* $\mathsf{C} : \{*\} \times X \times Y \to *$, $(\mathbb{P}\mathbb{T}, \mathsf{C}, \mathsf{G}, \mathsf{X})$ *is a see-do forecast exchangeable with respect to* $\mathsf{G}$.

*Proof.* For arbitrary $\rho : A \to A$, associated swap kernel $R^{\mathsf{D}}$ and $R^{\mathsf{X}}$:

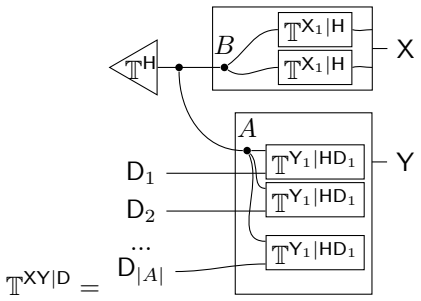
$$\begin{aligned}
\mathbb{P}\mathbb{T}R^{\mathsf{DX}} &= (\mathbb{P}R^{\mathsf{D}})\mathbb{T}R^{\mathsf{DX}} && (3.64)\\
&= \mathbb{P}(R^{\mathsf{D}}\mathbb{T}R^{\mathsf{DX}}) && (3.65)\\
&= \mathbb{P}\mathbb{T} && (3.66)
\end{aligned}$$

This is sufficient for exchangeability of $(\mathbb{P}\mathbb{T}, \mathsf{C}, \mathsf{G}, \mathsf{O})$ with respect to $\mathsf{G}$. □

**Theorem 3.2.13** (Representation of doubly exchangeable see-do forecasts)**.** *Given a see-do forecast* $(\mathbb{T}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ *with denumerable* $D$ *and standard measurable* $X$*,* $Y$*, the following statements are equivalent (given finite* $A \subset \mathbb{N}$*,* $B \subset \mathbb{N}$ *not necessarily finite):*

1. $(\mathbb{T}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ *is doubly exchangeable with respect to* $\mathsf{D} = \underline{\otimes}_{i\in A}\mathsf{D}_i$, $\mathsf{X} = \underline{\otimes}_{i\in B}\mathsf{X}_i$, $\mathsf{Y} = \underline{\otimes}_{i\in A}\mathsf{Y}_i$

2. *There exists a set* $H$, $\mathbb{T}^{\mathsf{H}} \in \Delta(\mathcal{H})$ *and Markov kernels* $\mathbb{T}^{\mathsf{X}_1|\mathsf{H}}$ *and* $\mathbb{T}^{\mathsf{Y}_1|\mathsf{H}\mathsf{D}_1}$ *such that*




$$\mathbb{T}^{\mathsf{X}\mathsf{Y}|\mathsf{D}} = \tag{3.67}$$

3. *There exists a set* $H$, $\mu \in \Delta(\mathcal{H})$ *and Markov kernels* $\mathbb{T}^{\mathsf{H}}$, $\mathbb{T}^{\mathsf{X}_1|\mathsf{H}}$ *and* $\mathbb{T}^{\mathsf{Y}_1|\mathsf{H}\mathsf{D}_1}$ *such that for all* $\mathbf{d}_A \in D$, $\{J_i \in \mathcal{X}_1 | i \in B\}$, $\{K_i \in \mathcal{Y}_1 | i \in A\}$*:*

$$\mathbb{T}^{\mathsf{X}\mathsf{Y}|\mathsf{D}}_{\mathbf{d}_A}((\times_{i\in B} J_i) \times (\times_{j\in A} K_j)) = \int_H \prod_{i\in B} \mathbb{T}^{\mathsf{X}_1|\mathsf{H}}_h(J_i) \prod_{i\in A} \mathbb{T}^{\mathsf{Y}_1|\mathsf{H}\mathsf{D}_1}_{h,d_i}(K_i) d\mathbb{T}^{\mathsf{H}}(h) \tag{3.68}$$

*Proof.* (2) and (3) are string and integral notation for the same statement.

(1) $\implies$ (2):

Define $\mathbb{P} \in \Delta({\mathcal{D}_\infty}^{\mathbb{N}})$ such that $\mathbb{P} = \otimes_{i\in\mathbb{N}}\mathbb{P}^{\mathsf{D}_1}$ for some strictly positive $\mathbb{P}^{\mathsf{D}_1}$ (recall that $D$ and hence $D_1$ is denumerable). $\mathbb{P}$ is exchangeable and independent and identically distributed. Consider some infinite doubly exchangeable extension $\mathbb{T}'$ of $\mathbb{T}$. Then $\mathbb{P}\mathbb{T}'$ is exchangeable with respect to $\mathsf{D}\mathsf{Y}' := \underline{\otimes}_{i\in\mathbb{N}}\mathsf{D}'_i\underline{\otimes}\mathsf{Y}'_i$ (Lemma 3.2.12) and exchangeable with respect to $\mathsf{X}' = \underline{\otimes}_{i\in\mathbb{N}}\mathsf{X}'_i$ as $\mathsf{X}$ is independent of $\mathsf{D}$.

By Lemma 3.2.10 we have $\mathsf{Z}' := f \circ \mathsf{X}'$ for some $f$ such that

1. $\mathsf{X}'_i \perp\!\!\!\perp_{\mathbb{P}\mathbb{T}'} \mathsf{X}'_{\mathbb{N}\setminus\{i\}} | \mathsf{Z}'$ for all $i \in A$

2. $(\mathbb{P}\mathbb{T}')^{\mathsf{X}'_i|\mathsf{Z}''} = (\mathbb{P}\mathbb{T}'')^{\mathsf{X}'_j|\mathsf{Z}''}$ for all $i, j \in A$

3. $\mathsf{D}\underline{\otimes}\mathsf{Y}' \perp\!\!\!\perp_{\mathbb{P}\mathbb{T}'} \mathsf{X}'|\mathsf{Z}'$

Applying Lemma 3.2.10 to $\mathsf{D}\mathsf{Y}'$, and noting that $\mathsf{D}'\underline{\otimes}\mathsf{Y}'$ is an invertible function of $\mathsf{D}\mathsf{Y}'$, we have $\mathsf{W}' = g \circ \mathsf{D}\mathsf{Y}'$ for some $g$ such that

4. for all $i \in \mathbb{N}$, $\mathsf{D}_i\underline{\otimes}\mathsf{Y}'_i \perp\!\!\!\perp_{\mathbb{P}\mathbb{T}'} \mathsf{D}'_{\mathbb{N}\setminus\{i\}}\mathsf{Y}'_{\mathbb{N}\setminus\{i\}} | \mathsf{W}'$

5. $(\mathbb{P}\mathbb{T}')^{\mathsf{Y}'_i|\mathsf{D}'_i} = (\mathbb{P}\mathbb{T}')^{\mathsf{Y}'_j|\mathsf{D}'_j}$ for all $i, j \in \mathbb{N}$

6. $\mathsf{X}' \perp\!\!\!\perp_{\mathbb{P}\mathbb{T}'} \mathsf{D}'\underline{\otimes}\mathsf{Y}' | \mathsf{W}'$

Because $\mathsf{W}' \underline{\otimes} \mathsf{D}'$ is a function of $\mathsf{DY}'$, we also have $\mathsf{X}' \perp\!\!\!\perp_{\mathbb{PT}'} \mathsf{W}' \underline{\otimes} \mathsf{D}' | \mathsf{Z}'$ by property 6.

$\mathsf{D}'$ is also a function of $\mathsf{DY}'$ and $\mathsf{Z}'$ is a function of $\mathsf{X}'$ so $\mathsf{D}' \perp\!\!\!\perp_{\mathbb{PT}'} \mathsf{Z}' | \mathsf{W}'$, also by property 6. Because $\mathbb{P}$ is independent and identically distributed, $\mathsf{D}' \perp\!\!\!\perp_{\mathbb{PT}'} \mathsf{W}'$, so $\mathsf{D}' \perp\!\!\!\perp_{\mathbb{PT}'} \mathsf{Z}' \underline{\otimes} \mathsf{W}'$

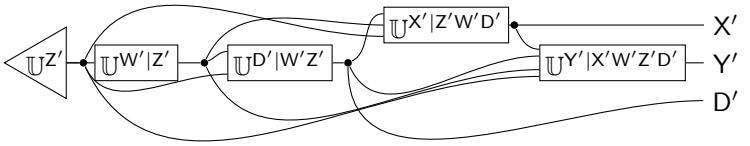
Becuase $\mathsf{Z}'$ is a function of $\mathsf{X}'$, $\mathsf{Y}' \perp\!\!\!\perp_{\mathbb{PT}'} \mathsf{Z}' | \mathsf{D}' \underline{\otimes} \mathsf{X}' \underline{\otimes} \mathsf{W}'$.

Applying weak union and symmetry to property 6 we have $\mathsf{Y}' \perp\!\!\!\perp \mathsf{X}' | \mathsf{W}' \underline{\otimes} \mathsf{D}'$, which combined with the above gives $\mathsf{Y}' \perp\!\!\!\perp \mathsf{X}' \underline{\otimes} \mathsf{Z}' | \mathsf{W}' \underline{\otimes} \mathsf{D}'$ by contraction.

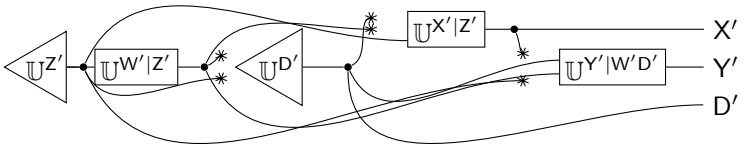
In summary, we will use the following conditional independences:

- $\mathsf{X}' \perp\!\!\!\perp_{\mathbb{PT}'} \mathsf{W}' \underline{\otimes} \mathsf{D}' | \mathsf{Z}'$
- $\mathsf{D}' \perp\!\!\!\perp_{\mathbb{PT}'} \mathsf{Z}' \underline{\otimes} \mathsf{W}'$
- $\mathsf{Y}' \perp\!\!\!\perp_{\mathbb{PT}'} \mathsf{Z}' | \mathsf{D}' \underline{\otimes} \mathsf{X}' \underline{\otimes} \mathsf{W}'$
- $\mathsf{Y}' \perp\!\!\!\perp \mathsf{X}' \underline{\otimes} \mathsf{Z}' | \mathsf{W}' \underline{\otimes} \mathsf{D}'$

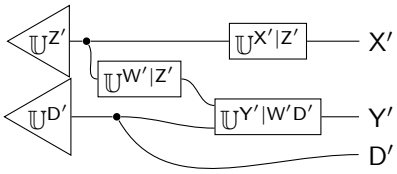
Let $\mathbb{U} := \mathbb{PT}'$. By Lemma 2.0.21, and




$(\mathbb{PT}')^{\mathsf{X}'\mathsf{Y}'\mathsf{D}'} =$ (3.69)



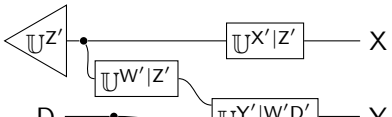

$=$ (3.70)



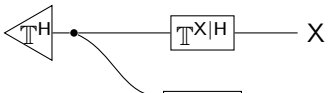

$=$ (3.71)

By mutual indepenence of $\mathsf{Y}'_i \underline{\otimes} \mathsf{D}'_i$'s', we have in particular $\mathsf{X}'_A \underline{\otimes} \mathsf{Y}'_A \underline{\otimes} \mathsf{D}'_A \perp\!\!\!\perp_{\mathbb{U}} \mathsf{X}'_{A^C} \underline{\otimes} \mathsf{Y}'_{A^C} \underline{\otimes} \mathsf{D}'_{A^C}$. Therefore $\mathbb{U}^{\mathsf{X}'_A \mathsf{Y}'_A \mathsf{D}'_A} = (\mathbb{P}^{\mathsf{D}_A} \mathbb{T})$. Furthermore, $\mathbb{P}^{\mathsf{D}_A}$ is positive by assumption, so by Lemma 2.0.30:

$$\mathbb{T} = \mathbb{U}^{\mathsf{X}'_A \mathsf{Y}'_A | \mathsf{D}'_A} \tag{3.72}$$



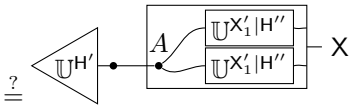

$$= \tag{3.73}$$




$$\stackrel{def}{=} \tag{3.74}$$

We still need to show

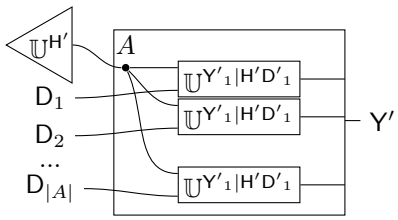
$$\mathbb{T}^{\mathsf{X}} = \mathbb{U}^{\mathsf{X}'_A} \tag{3.75}$$




$$\stackrel{?}{=} \tag{3.76}$$

and

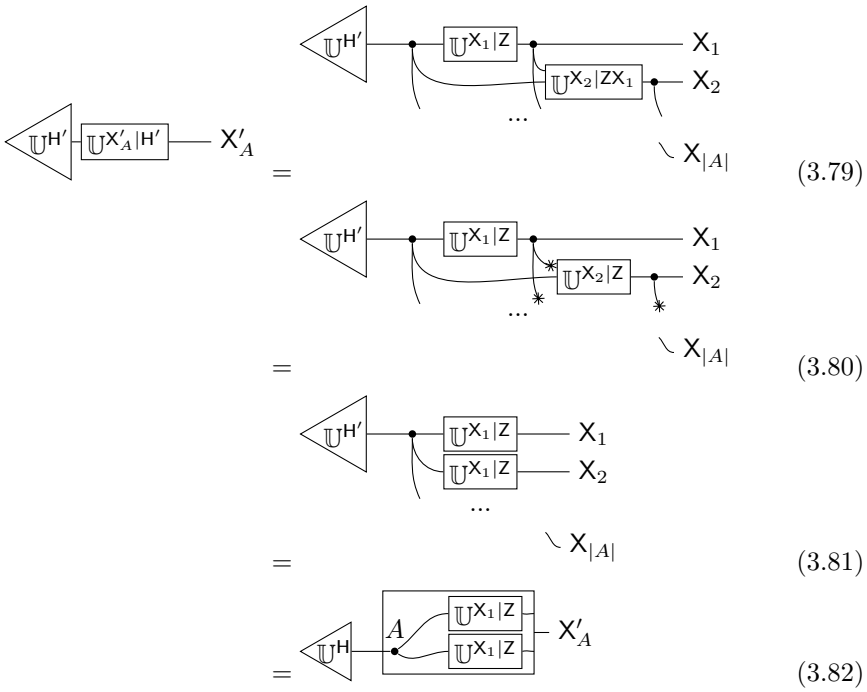
$$\mathbb{T}^{\mathsf{Y}|\mathsf{D}} = \mathbb{U}^{\mathsf{Y}'_A | \mathsf{D}'_A} \tag{3.77}$$




$$\stackrel{?}{=} \tag{3.78}$$

Where $\mathsf{H}' = \mathsf{Z}' \underline{\otimes} \mathsf{W}'$.

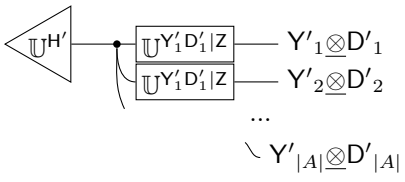
The first follows straightforwardly from the mutual independence of the $\mathsf{X}'_i$'s. The following diagram uses "..." informally to indicate a missing section of diagram that continues "as you'd expect". Applying Lemma 2.0.21, we have




As desired.

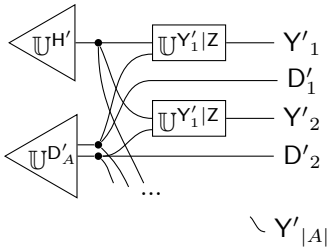
For $\mathbb{U}^{\mathsf{Y}'_A|\mathsf{D}'_A}$, define the "interleaved" random variable $\mathsf{YD}'_A = \underline{\otimes}_{i\in A}\mathsf{Y}'_i\underline{\otimes}\mathsf{D}'_i$. Note that this implies that $\mathbb{U}^{\mathsf{Y}'_A\mathsf{D}'_A} = \mathbb{U}^{\mathsf{YD}'_A}\rho$ where $\rho$ is a swap kernel that moves all the $\mathsf{D}'_i$s below the $\mathsf{Y}'_i$s. Then, applying what we just showed but substituting $\mathsf{YD}'_A$ for $\mathsf{X}'_A$,

$$\mathbb{U}^{\mathsf{YD'}_A} =$$




(3.83)

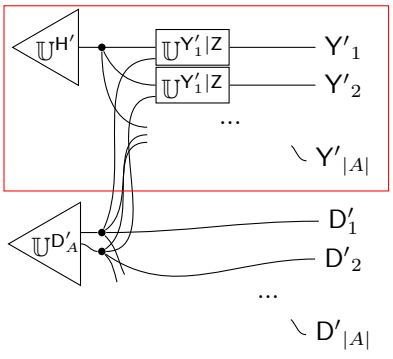
$$=$$




(3.84)

$$\implies \mathbb{U}^{\mathsf{Y'}_A\mathsf{D'}_A} = \mathbb{U}^{\mathsf{YD'}_A}\rho \tag{3.85}$$

$$=$$




(3.86)

The red boxed kernel is the disintegration $\mathbb{U}^{\mathsf{Y'}_A|\mathsf{D'}_A}$ and (notational difficulties aside) is equal to the kernel in Equation 3.78.

This completes the proof that (1) $\implies$ (2).

(2) $\implies$ (1):

We will use integral notation. For all sets of events $\{J_i \in \mathcal{X}_1\}_B$, $\{K_i \in \mathcal{Y}_1\}_A$, $\mathbf{d}_A \in D$:

$$\mathbb{T}^{\mathsf{XY}|\mathsf{D}}_{\mathbf{d}_A}((\times_{i\in B} J_i) \times (\times_{j\in A} K_j)) = \int_H \prod_{i\in B} \mathbb{T}^{\mathsf{X}_1|\mathsf{H}}_h(J_i) \prod_{i\in A} \mathbb{T}^{\mathsf{Y}_1|\mathsf{HD}_1}_{h,d_i}(K_i) d\mathbb{T}^{\mathsf{H}}(h) \tag{3.87}$$

For all $\{J_i \in \mathcal{X}_1\}_{\mathbb{N}}$, $\{K_i \in \mathcal{Y}_1\}_{\mathbb{N}}$, $\mathbf{d} \in D^{\mathbb{N}}$ define

$$\mathbb{T}_{\mathbf{d}}'^{\mathsf{X}'\mathsf{Y}'|\mathsf{D}'}((\times_{i\in\mathbb{N}}J_i)\times(\times_{j\in\mathbb{N}}K_j)) = \int_H \prod_{i\in\mathbb{N}} \mathbb{T}_h^{\mathsf{X}_1|\mathsf{H}}(J_i) \prod_{i\in\mathbb{N}} \mathbb{T}_{h,d_i}^{\mathsf{Y}_1|\mathsf{H}\mathsf{D}_1}(K_i) d\mathbb{T}^{\mathsf{H}}(h) \tag{3.88}$$

Marginalising over $\mathbb{N}\setminus A$ and $\mathbb{N}\setminus B$ is equivalent to choosing $K_i = Y_1$ for $i \notin A$ and $J_i = X_1$ for $i \notin B$. Then

$$\mathbb{T}_{\mathbf{d}}'^{\mathsf{X}'_B\mathsf{Y}'_A|\mathsf{D}'}((\times_{i\in B}J_i\times X_1^{\mathbb{N}\setminus B})\times(\times_{j\in A}K_j\times Y_1^{\mathbb{N}\setminus A})) = \int_H \prod_{i\in B} \mathbb{T}_h^{\mathsf{X}_1|\mathsf{H}}(J_i)[\mathbb{T}_h^{\mathsf{X}_1|\mathsf{H}}(X_1)]^{\mathbb{N}\setminus B} \prod_{i\in A} \mathbb{T}_{h,d_i}^{\mathsf{Y}_1|\mathsf{H}\mathsf{D}_1}(K_i)[\mathbb{T}_h^{\mathsf{Y}_1|\mathsf{H}}(Y_1)]^{\mathbb{N}\setminus A} d \tag{3.89}$$

$$= \int_H \prod_{i\in B} \mathbb{T}_h^{\mathsf{X}_1|\mathsf{H}}(J_i) \prod_{i\in A} \mathbb{T}_{h,d_i}^{\mathsf{Y}_1|\mathsf{H}\mathsf{D}_1}(K_i) d\mathbb{T}^{\mathsf{H}}(h) \tag{3.90}$$

$$= \mathbb{T}_{\mathbf{d}_A}^{\mathsf{X}\mathsf{Y}|\mathsf{D}}((\times_{i\in B}J_i)\times(\times_{j\in A}K_j)) \tag{3.91}$$

Thus $\mathbb{T}'$ is an extension of $\mathbb{T}$. Furthermore, for any swaps $\rho_r$, $\rho_s$ with associated kernels $R^{\mathsf{X}'}$, $S^{\mathsf{D}'}$,$S^{\mathsf{Y}'}$:

$$(S^{\mathsf{D}'\mathsf{Y}'}\mathbb{T}'(R^{\mathsf{X}'}\otimes S^{\mathsf{Y}'}))_{\mathbf{d}}^{\mathsf{X}'\mathsf{Y}'|\mathsf{D}'}((\times_{i\in\mathbb{N}}J_i)\times(\times_{j\in\mathbb{N}}K_j)) = \int_H \prod_{i\in\mathbb{N}} \mathbb{T}_h^{\mathsf{X}_1|\mathsf{H}}(J_{\rho_r(i)}) \prod_{i\in\mathbb{N}} \mathbb{T}_{h,d_{\rho_s(i)}}^{\mathsf{Y}_1|\mathsf{H}\mathsf{D}_1}(K_{\rho_s(i)}) d\mathbb{T}^{\mathsf{H}}(h) \tag{3.92}$$

$$= \int_H \prod_{i\in\rho_r(\mathbb{N})} \mathbb{T}_h^{\mathsf{X}_1|\mathsf{H}}(J_i) \prod_{i\in\rho_s(\mathbb{N})} \mathbb{T}_{h,d_i}^{\mathsf{Y}_1|\mathsf{H}\mathsf{D}_1}(K_i) d\mathbb{T}^{\mathsf{H}}(h) \tag{3.93}$$

$$= \mathbb{T}_{\mathbf{d}}'^{\mathsf{X}'\mathsf{Y}'|\mathsf{D}'}((\times_{i\in\mathbb{N}}J_i)\times(\times_{j\in\mathbb{N}}K_j)) \tag{3.94}$$

Therefore $\mathbb{T}$ is infinitely doubly exchangeably extendable. □

# Chapter 4

# Statistical Decision Theory

I think I've got a good idea for a result that is relevant to the discussion below

*Proposition:* If you have a decision problem of the following form:

- You get an observation $x$ and may return some mixtures of decisions in $\Delta(\mathcal{D})$; that is, you can choose some function $x \to \Delta(\mathcal{D})$
- For any function $Q : x \to \Delta(\mathcal{D})$ we have a forecast of the observations, decisions and consequences, and it is appropriate to model the forecast with some probability over observations decisions and consequences $\mathbb{P}_Q \in \Delta(\mathcal{X} \otimes \mathcal{D} \otimes \mathcal{Y})$
  - You can probably make this a collection of probabilities/Markov kernel, it's just simpler to consider one probability for this outline
- For all $Q, R : x \to \Delta(\mathcal{D})$, $\mathbb{P}_Q^{\mathsf{X}} = \mathbb{P}_R^{\mathsf{X}}$, $\mathbb{P}_Q^{\mathsf{Y}|\mathsf{XD}} = \mathbb{P}_R^{\mathsf{Y}|\mathsf{XD}}$ and $\mathbb{P}_Q^{\mathsf{D}|\mathsf{X}} = Q$
  - That is: I should expect the same observations whatever function I end up choosing and I expect the same consequences holding the observations and decision fixed (even more informally: it doesn't matter how I choose decisions provided I end up choosing the same one)
  - Finally, the decision function we choose is the relation between $\mathsf{X}$ and $\mathsf{D}$
- There exists some $Q$ such that $\mathbb{P}_Q$ has full support

Then there is a unique see-do forecast $\mathbb{T}$ such that for every $Q : x \to \Delta(\mathcal{D})$, $\mathbb{T}$ "intertwined" with $Q$ is equal to $\mathbb{P}_Q$ (intertwining being a well-defined operation I haven't written down).

*Proof sketch:* Jacobs et al. (2019) introduces the notion of a "2-comb" which is a Markov kernel with two inputs, two outputs and one of the outputs is

independent of one of the inputs. This is essentially equivalent to a see-do model (Definition 3.1.2) - "essentially" because we need to give wires names to get a see-do model, but the independence condition means that there is always a unique way to do this.

Jacobs et. al. prove a *comb disintegration* theorem: given any probability with 3 wires and full support, there is a unique 2-comb and a Markov kernel that can be "intertwined" to give the original probability. Thus we can get a unique 2-comb $\mathsf{V}$ for some $Q$ where $\mathbb{P}_Q$.

Further, we can get a collection of 2-combs $\mathsf{U}, \mathsf{W}$,... for $R, S : x \to \Delta(\mathcal{D})$ where $\mathbb{P}_R$ may or may not have full support.

If we have two 2-combs $\mathbb{U}$ and $\mathbb{V}$ such that $\mathbb{U}^{\mathsf{X}|\mathsf{H}} = \mathbb{V}^{\mathsf{X}|\mathsf{H}}$ and $\mathbb{U}^{\mathsf{Y}|\mathsf{XHD}} = \mathbb{V}^{\mathsf{Y}|\mathsf{XHD}}$ then $\mathbb{U} = \mathbb{V}$ (applying Lemma 2.0.21). Thus $\mathbb{U}$ is a unique 2-comb that every probability $\mathbb{P}_Q$, $\mathbb{P}_R$ etc. can be disintegrated to.

*Comment:* This is a "soft representation theorem". While the theorems discussed below try to establish the need for expected utility and probability from constraints on rational preferences or beliefs, this one says if you're already happy to use probability then you can represent your knowledge with a see-do forecast/see-do model

Also, it ties in with von Neumann-Morgenstern: we have assumed at the outset that we are choosing between $\mathbb{P}_Q$, $\mathbb{P}_R$ etc., which are lotteries (albeit non-standard ones) in the language of vNM.

It might also tie in with Walley - if we have a *coherent prevision* for each decision function, then this can be uniquely represented with a set of probailities.

End of the sketch of the result

## 4.1 Representing prior knowledge in decision problems

We introduced see-do models in Chapter 3 and gave the random variables suggestive names: $\mathsf{D}$ is the "choices", $\mathsf{X}$ is the "observations", $\mathsf{H}$ is the "hypothesis" and $\mathsf{Y}$ is the "consequences". However, as we discussed, the actual interpretations of these variables depend on what exactly the model is being used for. One use case is making decisions.

Consider an idealised decision problem where a set of observations will be given which may take values in $X$, a utility function $u : Y \to \mathbb{R}$ is given and, after viewing the observations, one must select one decision from a known set $D$ and anticipates some element of $Y$ to occur as a consequence. The utility function measures which elements of $Y$ are preferred, where $u(y) > u(y')$ means $y$ is preferred to $y'$, and one aims to select a decision that results in an element of $Y$ that is preferred to other elements of $Y$ according to $u$.

In order to solve this kind of problem, we need some additional knowledge: we cannot choose a decision in $D$ on the basis of preferences over $Y$ unless we know decisions have some effect on which elements of $Y$ are likely to occur. In many, it is unlikely that we could ever know that each particular decision will

force a particular element of $Y$ to occur, so we need some means of representing the consequences of decisions with uncertainty. We might use probability to represent all relevant uncertainty, and so each decision is associated with a particular probability distribution in $\Delta(\mathcal{Y})$. Alternatively, we could entertain two types of uncertainty: we continue to suppose each decision is associated with a probability distribution in $\Delta(\mathcal{Y})$ but we have some uncertainty about which map $D \to \Delta(\mathcal{Y})$ is appropriate and we are not willing to represent this uncertainty probabilistically, and choose a *nonempty set* of maps $D \to \Delta(\mathcal{Y})$ to represent our full uncertainty.

If our uncertainty is represented by a single map $D \to \Delta(\mathcal{Y})$ then we have a do-forecast (Definition 3.2.1. If our uncertainty is represented by a set of maps, then we can assign each map an index $h \in H$ for some set of hypotheses $H$ such that $h(d) \in \Delta(\mathcal{Y})$. Assuming measurability, then $\mathbb{T} : H \times D \to \Delta(\mathcal{Y})$ given by $\mathbb{T} : (h, d) \mapsto h(d)$ is a two-player statistical model (Definition 3.1.1).

We will often also wish to make use of the observations $X$ to improve our decision. If we are fortunate, the observations can reduce our uncertainty. We may wish to represent joint uncertainty as a single map to joint distributions $\mathbb{T} : D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ and "make use of" an observation $x \in X$ by conditioning $\mathbb{T}$ on the event $\mathbb{1}_{\mathsf{X}=x}$. Alternatively, we could choose a set of such maps and "make use of" observations by conditioning each map on them. If we add the assumption $\mathsf{X} \perp\!\!\!\perp \mathsf{D}$ (which will be explained in more detail in this chapter) for all such maps, in the case of a single map we have a see-do forecast (Definition 3.2.1) and in the case of sets of maps we have a see-do model (Definition 3.1.2).

Define conditioning chapter 2

### 4.1.1 How should knowledge be represented?

Decision problems require certain types of additional knowledge, and see-do models or forecasts can represent knowledge of this type. We can ask if and when it is a sound choice to use these kinds of models for this purpose. Our answer, for the time being, is to lean on prior work: we will show in in Chapter 5 that graphical causal models are a subset of see-do models and that two-player statistical models in conjunction with a particular set of assumptions defining "counterfactual random variables" yield standard potential outcomes models. Finally, in this chapter we will see how see-do models combined with utility yield statistical decision problems. As all of these are types of see-do model, or can be derived from see-do models along with additional provisions, it follows that if any of them are sometimes sound approaches for representing knowledge in decision problems then see-do models must be sound choices.

An alternative approach to answering this question is to define a set of axioms that we want our knowledge representation to satisfy and then show that these axioms are compatible with see-do models or even that these axioms imply see-do models. There is a body of literature of this genre and we will briefly survey results and comment on their relationship to see-do models. Most such "representation theorems" aim to show that a particular means of representing

knowledge is *necessary* given a number set of assumptions that purport to define "rational preferences" or "rational beliefs". The question of whether a given set of axioms are in fact required to call an agents preferences or beliefs rational is a difficult one, and we will not have anything to add to that discussion here – neither by commenting on axioms already proposed or by proposing our own. However, we are still interested in comparing see-do models with model types used in existing representation theorems.

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

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

This defintion is meant to correspond to the common sense idea of having preferences over some large 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 $a''$.

### 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" (which for our purposes are probability distributions in $\Delta(\mathcal{Y})$) can be represented with a utility function $u : Y \to \mathbb{R}$ unique up to affine transformation along with the principle of expected utility, which is, for lotteries $\mathbb{P}, \mathbb{P}' \in \Delta(\mathcal{Y})$ the rule that $\mathbb{E}_{\mathbb{P}}[u] > \mathbb{E}_{\mathbb{P}'}[u]$ is equivalent to the statement $\mathbb{P} \succ \mathbb{P}'$.

If we consider a set of decisions $D$ and a do-forecast $\mathbb{T}^{\mathsf{Y}|\mathsf{D}}$, then we can identify the set of lotteries with the evaluations of each decision under $\mathbb{T}$ – that is, the lotteries are $\{\mathbb{T}_d^{\mathsf{Y}|\mathsf{D}} | d \in D\}$. We can then define preferences over decisions: $d$ is strictly preferred to $d'$ if and only if $\mathbb{T}_d^{\mathsf{Y}|\mathsf{D}}$ is strictly preferred to $\mathbb{T}_{d'}^{\mathsf{Y}|\mathsf{D}}$. Then, if preferences over the set of lotteries satisfies the *vNM axioms*, there exists a utility function $u : Y \to \mathbb{R}$ such that

- $d \succ d'$ if and only if $\mathbb{T}_d^{\mathsf{Y}|\mathsf{D}} u > \mathbb{T}_{d'}^{\mathsf{Y}|\mathsf{D}} u$

Is this worth stating as a theorem?

### Savage's decision theory

The von Neumann-Morgenstern theorem shows that, if we take do-forecasts for granted and assuming the vNM axioms then we can represent preferences

between decisions using expected utility. We are more interested in when we should use do-forecasts or see-do forecasts or see-do models.

Savage (1954) proved a representation theorem for decision problems, which has more overlap with our original question - namely, what should we use to represent the knowledge we bring to a kind of idealised decision problem. Savage's decision problems featured *states* $H$ (which we can be identified with hypotheses in Definition 3.1.1), *acts* $D$ (which can be identified with decisions in 3.1.1) and *outcomes* $O$ (which can be identified with outcomes in Definition 3.1.1). Savage takes it as a given that we have a deterministic two player statistical model $(\mathbb{T}, \mathsf{H}, \mathsf{D}, \mathsf{O})$, and furthermore for *any* function $f : H \to O$ there exists some $d \in D$ such that $\mathbb{T}_{h,d} = \delta_{f(h)}$, and furthermore we have a preference relation on $D$. He then shows that if the *Savage axioms* hold for the preference relation on $D$ then there exists a unique probability measure $\mathbb{P} \in \Delta(\mathcal{H})$ and a utility $u : O \to \mathbb{R}$ unique up to affine transformation such that

$$(d \succ d') \iff (\mathbb{P} \otimes \delta_d)\mathbb{T}u > (\mathbb{P} \otimes \delta_{d'})\mathbb{T}u \tag{4.1}$$

The use of two player statistical models is baked into Savage's representation theorem as a basic assumption. However, he also shows that if the Savage axioms hold for preferences among outcomes, then it is possible to define a unique do-forecast

$$\mathbb{F} := (\mathbb{P} \otimes \mathrm{Id}_D)\mathbb{T} \tag{4.2}$$

(using the existing definitions for $\mathsf{D}$ and $\mathsf{O}$) such that

$$(d \succ d') \iff \mathbb{F}_d \mathbb{T}u > \mathbb{F}_d u \tag{4.3}$$

Is this worth stating as a theorem?

**Jeffrey's decision theory**

The approach to decision making set out by Jeffrey (1965) and Bolker (1966) differs from Savage's approach in a number of ways. Firstly, the representation theorem proved by Bolker does not distinguish between hypotheses, outcomes and decisions. It assumes only an algebra of outcomes $(O, \mathcal{O})$ and a preference relation over these outcomes. If the preference relation satisfies the *Jeffrey axioms* then there exists a utility $u : O \to \mathbb{R}$ and a probability distribution $\mathbb{P} \in \Delta(\mathcal{O})$ which are non-unique such that for $A, B \in \mathcal{O}$ and finite partition $C_1, ..., C_n \in \mathcal{O}$ (the proof is given by Bolker (1966)):

$$(A \succ B) \iff (\mathbb{P}_1^{\mathsf{O}|\mathbb{1}_A} u > \mathbb{P}_1^{\mathsf{O}|\mathbb{1}_B} u) \tag{4.4}$$

A key feature to note here is that instead of comparing two decisions by evaluating one Markov kernel at two different points (as in our theory and Savage's), Jeffrey compares two events by comparing two different Markov kernels $\mathbb{P}^{\mathsf{O}|\mathbb{1}_A}$ and $\mathbb{P}^{\mathsf{O}|\mathbb{1}_B}$. In order to use such a model in a decision problem, Jeffrey suggests we can identify a subset of the events with the choices we can make: $D := \{D_i \in \mathcal{O} | i \in A\}$ for some set $A$, so we get one Markov kernel for each decision we might make.

Whether we should model decisions as a set $D$ and consequence maps as Markov kernels with $D$ as the domain or as a collection of events and consequence maps as a corresponding collection of Markov kernels, or something else is a somewhat subtle issue that I don't fully understand.

I think Jeffrey goes too far in saying decisions are events. Consider the problem of deciding whether or not to order a drink, and suppose that I choose to consider a set of options $D = \{d_1 =$ "order a drink and make sure that drink is water", $d_2 =$ "order a drink", $d_3 =$ "don't order a drink"$\}$. If I choose $d_1$, then it will certainly be true that I will order a drink and also true that I will order a drink of water, and in general an event "A and B" implies an event "A". However, $d_1$ and $d_2$ are distinct – if I particularly want to drink water then I will choose $d_1$ and not $d_2$. This feature of decision making is reflected in our idealised decision problem at the start where we may make a single decision from our set of available choices.

At the same time, it does seem that in the same way that events leave many features of the outcome underspecified, when we describe choices we very often leave features of what we precisely intend to do underspecified. Choices aren't necessarily underspecified: imagine a reinforcement learner in a simulated discrete time environment. At each timestep this learner takes its history so far and outputs some action and, if we examine the code governing how it interacts with its environment we can probably deduce *all* the actions available to it at each timestep. In less controlled contexts, which is really most contexts we are familiar with, if we want to come up with a list of decisions we could make then all the decisions in this list are likely to leave a large number of things that we could in principle decide on underspecified. In the example above, it seems I could decide to:

- Order a drink
- Order a glass of water
- Order a glass of water and say please and thankyou when I do
- Order a glass of water, say please and thankyou, carefully avoid scratching my itchy ear

And so forth. I could go on adding details for a very long time without exhausting the set of things I could in principle decide to do. I usually would not want to consider all these "in principle" choices I have available. Many of the details are more or less irrelevant and not worth spending the time to think about. Furthermore, even if I do consider all choices I have available in principle, it seems plausible that a less specific decision could be preferable. For example

I might expect the "automatic execution" of some things things I didn't fully specify to be closer to optimal than the best guess I have of how to specify them (consider managing a competent employee; one might get better results by being less specific in requests).

I think the question of how one could structure the decisions set $D$ is an interesting one. I have so far treated it as "just a set" with no particular additional structure. Jeffrey suggests that decisions are a collection of events, and decisions do seem to have some event-like features, like the fact that they specify incompletely what I will do next and the fact that different specifications can apparently be joined with "and". However, the proposition that decisions are identical to events does not seem quite right because, as in the example above, deciding to do "A and B" does not imply deciding to do "A".

I think this is also potentially an important question. In Chapter 6 I discuss the assumption of *imitability*, that it is within a decision maker's power to reproduce the observed data. Such an assumption, if it holds, licences a number of inferences from data to consequences. It may often be a plausible assumption considering the full set of decisions $D^*$ that a decision maker could make *in principle*, but may be much less often plausible when considering the restricted set of decisions $D$ a decision maker is actually considering. Understanding how these may relate to one another may help to better understand assumptions like imitability.

Which has only just occurred to me

#### Causal decision theory

Statistical decision problems were introduced by Wald (1950). A statistical decision problem posits a set of hypotheses $H$ and observations $X$ and the two are related by a Markov kernel $\mathbb{M} : H \to \Delta(\mathcal{X})$ (which we can recognise as a *statistical model* from the previous chapter). In addition, a statistical decision problem involves a set of decisions $D$ and a loss $l : H \times D \to [0, \infty)$. Defining a Markov kernel $\mathbb{T} := (\mathbb{M} \otimes \mathrm{Id}_D) \underline{\otimes} \mathbb{F}_l$ by "joining" the statistical model and the loss and defining the consequences of a decision under a particular hypothesis to be, with probability 1, the loss induced by that decision in that hypothesis, we actually have a see-do model $(\mathbb{T}, \mathsf{H}, \mathsf{X}, \mathsf{Y})$. This model has the special properties that the consequences $\mathsf{Y}$ take values in $[0, \infty)$ and $\mathbb{T}^{\mathsf{Y}|\mathsf{DH}}$ is deterministic. The relationship between statistical decision problems and see-do models will be discussed further below – here, it is enough that the representation of a statistical decision problem is a special case of a see-do model.

### 4.1.2 Decision functions

- Define decision functions
- Choosing a particular mixed decision vs choosing a mixture of decisions; disintegrations exist both ways

### 4.1.3 Risk

- Define risk
- Define risk set

### 4.1.4 Reachable consequences

- Reachable consequences $\sim$ risk set without a utility

### 4.1.5 Decision rules

- Need a rule for selecting a decision function
- Admissibility
- Maximise EU w/a Bayes forecast
- Minimax
- Complete class theorem

### 4.1.6 Comparison of experiments and actuators

- Comparison of experiments
- Comparison of actuators
- Limiting cases: no information and no influence (thus: some information and some influence is necessary for nontrivial problem)

### 4.1.7 Equivalence of see-do models

- Definition of equivalence via reachable set
- Defintion + examples: decomposability
- Thm: an indecomposable see-do model has an equivalent decomposable model

## 4.2 Scraps to be moved into skeleton above

Currently a disorganised cut and paste

### 4.2.1 Decomposability

Decomposability is a property of see-do models that is relevant to the distinction between counterfacutal and regular models. As we will show, many causal problems allow the use of decomposable see-do models. However, certain types of counterfactual problem do not.

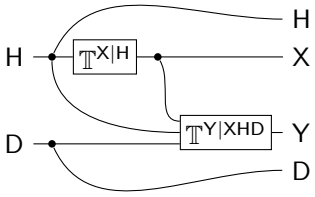
**Definition 4.2.1** (decomposability)**.** A see-do model $(\mathbb{T}, \mathsf{H}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ is *decomposable* iff $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{X}|\mathsf{DH}$. That is, if the consequence is independent of the observations given the hypothesis and the choice.

Decomposable see-do models can be represented as a pair $(\mathbb{B}, \mathbb{C})$ where $\mathbb{B}$ is a one-player statistical model we call the *observation model* and $\mathbb{C}$ is a two-player statistical model we call the *consequence model* (Corollary 4.2.3. Most models in the causal inference literature are decomposable – if the observed data can tell us nothing useful beyond the distribution of observations, then we have a decomposable model.

**Theorem 4.2.2** (Observation and Consequence models)**.** *Any see-do model* $(\mathbb{T}, \mathsf{H}, \mathsf{O}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ *can be uniquely represented by the following pair of Markov kernels:*

- *The* observation model $\mathbb{T}^{\mathsf{X}|\mathsf{H}}$
- *The* context-sensitive consequence model $\mathbb{T}^{\mathsf{Y}|\mathsf{XHD}}$

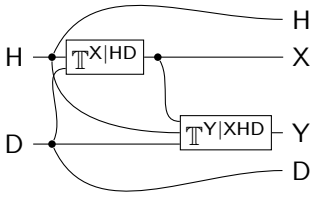
*Furthermore*




$\mathbb{T} =$ (4.5)

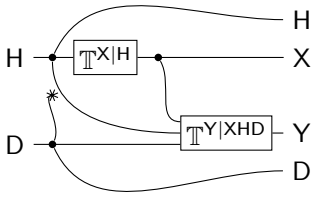
Maybe moves proofs out of main text

*Proof.* By Lemma 2.0.21,

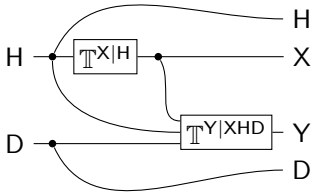



$\mathbb{T} =$ (4.6)

By the assumption $\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$ and version 2 of conditional independence from Theorem 2.0.33,

$\mathbb{T} =$ 

(4.7)

$=$ 

(4.8)

□

**Corollary 4.2.3.** *A decomposable see-do model $\mathbb{T} : H \times D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ can be uniquely represented by*

- *The* observation model $\mathbb{T}^{\mathsf{X}|\mathsf{H}}$
- *The* consequence model $\mathbb{T}^{\mathsf{Y}|\mathsf{HD}}$

*Proof.* Because $\mathbb{T}$ is decomposable, $\mathbb{T}^{\mathsf{Y}|\mathsf{XHD}} = *_X \otimes \mathbb{T}^{\mathsf{Y}|\mathsf{HD}}$. Then by Lemma 2.0.21 we have a unique representation of $\mathbb{T}$. □

### Examples of decomposable and indecomposable see-do models

Recall the previous example: suppose we are betting on the outcome of the flip of a possibly biased coin with payout 1 for a correct guess and 0 for an incorrect guess, and we are given $N$ previous flips of the coin to inspect. This situation can be modeled by a decomposable see-do model. Define $\mathbb{B} : (0,1) \to \Delta(\{0,1\})$ by $\mathbb{B} : \mathsf{H} \mapsto \text{Bernoulli}(\mathsf{H})$. Then define ${}^1\mathbb{T}$ by:

- $D = \{0,1\}$
- $X = \{0,1\}^N$
- $Y = \{0,1\}$
- $H = (0,1)$

- ${}^1\mathbb{B} : \curlyvee^N \mathbb{B}$
- ${}^1\mathbb{C} : (h, d) \mapsto \text{Bernoulli}(1 - |d - h|)$

In this model, the chance $\mathsf{H}$ of the coin landing on heads is as much as we can hope to know about how our bet will work out.

Suppose instead that in addition to the $N$ prior flips, we manage to look at the outcome of the flip on which we will bet. In this case, the situation can be modeled by the following indecomposable see-do model ${}^2\mathbb{T}$:

- $D = \{0, 1\}$
- $X = \{0, 1\}^{N+1}$
- $Y = \{0, 1\}$
- $H = (0, 1)$
- ${}^2\mathbb{T}^{\mathsf{X}|\mathsf{H}} : \curlyvee^{N+1}\mathbb{B}$
- ${}^2\mathbb{T}^{\mathsf{Y}|\mathsf{XHD}} : (h, \mathbf{x}, d) \mapsto \delta_{1-|d-x_{N+1}|}$

In this case, even if we are told the value of $\mathsf{H}$, we still benefit from using the observed data when making our decision.

It is possible to model the second situation with a decomposable model by including the result of the $N + 1$th flip in the hypothesis. Define the new hypothesis space $H' = H \times \{0, 1\}$ and let $\mathsf{H}_0$ be the projection to the old hypothesis space $H$. Define ${}^3\mathbb{T}$ by:

- $D = \{0, 1\}$
- $X = \{0, 1\}^{N+1}$
- $Y = \{0, 1\}$
- $H' = (0, 1) \times \{0, 1\}$
- ${}^3\mathbb{B} : (\curlyvee^N \mathbb{B} \otimes \delta_{x_{N+1}}$
- ${}^3\mathbb{C} : (h, x_{N+1}, d) \mapsto \delta_{1-|d-x_{N+1}|}$

However, ${}^2\mathbb{T}^{\mathsf{X}_{N+1}|\mathsf{H}} = \mathbb{B}$ while ${}^3\mathbb{T}^{\mathsf{X}_{N+1}|\mathsf{H}_0}$ is undefined, so ${}^3\mathbb{T}$ is a substantially different model to ${}^2\mathbb{T}$.

If an indecomposable see-do model is employed in a *decision problem* it is possible to create an equivalent decision problem with a decomposable model as I will show later. Some counterfactual problems cannot be formulated as decision problems, and indecomposability is a property of the types of counterfacutal model proposed by Pearl (2009), but not to my knowledge of any causal models used in a "decision like context".

### 4.2.2 Causal questions and decision functions

Pearl and Mackenzie (2018) has proposed three types of causal question:

1. Association: How are $\mathsf{W}$ and $\mathsf{Z}$ related? How would observing $\mathsf{W}$ change my beliefs about $\mathsf{Z}$?
2. Intervention: What would happen if I do ... ? How can I make ... happen?
3. Counterfactual: What if I had done ... instead of what I actually did?

*Causal decision problems* are, roughly speaking, "interventional" problems. In English, a causal decision problem roughly asks

> Given that I have data $\mathsf{X}$ and I know which values of $\mathsf{Y}$ I would like to see and some knowledge about how the world works, which of my available choices $D$ should I select?

This type of question presupposes somewhat more than Pearl's prototypical interventional questions. First, it supposes that we have *preferences* over the values that $\mathsf{Y}$ might take, which we need not have to answer the question "What would happen if I do ...?". Secondly, and crucially to our theory, causal decision problem suppose that we are given data and a set of choices.

We will return to the question of preferences. For now, we will focus on the idea that a causal decision problem is about selecting a choice given data. That is, however the selection is made, the answer to a causal decision problem is always a *decision function* $\mathbb{D} : X \to \Delta(\mathcal{D})$.

A property that will be of interest when considering counterfactual models is *decomposability*. A see-do model

#### Example

Suppose we are betting on the outcome of the flip of a possibly biased coin with payout 1 for a correct guess and 0 for an incorrect guess, and we are given $N$ previous flips of the coin to inspect. This situation can be modeled by a decomposable see-do model. Define $\mathbb{B} : (0,1) \to \Delta(\{0,1\})$ by $\mathbb{B} : \mathsf{H} \mapsto \text{Bernoulli}(\mathsf{H})$. Then define $\mathbb{T}$ by:

- Choice set: $D = \{0,1\}$
- Observation set: $X = \{0,1\}^N$
- Consequence set: $Y = \{0,1\}$
- Hypothesis set: $H = (0,1)$
- Observation map: $\mathbb{T}^{\mathsf{X}|\mathsf{H}} : \curlyvee^N \mathbb{B}$
- Consequence model: $\mathbb{T}^{\mathsf{Y}|\mathsf{DH}} : (h,d) \mapsto \text{Bernoulli}(1 - |d - h|)$

In this model, the chance $\mathsf{H}$ of the coin landing on heads is as much as we can hope to know about the success of our bet. $\mathsf{H}$ may be inferred from observation by some standard method, and

### Avoiding indecomposability with decision functions

Show that a decision problem with a indecomposable model induces an equivalent decision problem with a decomposable model with an expanded set of choices, subject to some conditions.

### Decision rules

See-do models encode the relationship between observed data and consequences of decisions. In order to actually make decisions, we also require preferences over consequences. We suppose that a *utility function* is given, and evaluate the desirability of consequences using *expected utility*. A see-do model along with a utility allows us to evaluate the desirability of *decisions rules* according to each hypothesis.

**Definition 4.2.4** (Utility function)**.** Given a See-Do Model $\mathbb{T} : \mathsf{H} \times D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$, a *utility function* $u$ is a measurable function $Y \to \mathbb{R}$.

**Definition 4.2.5** (Expected utility)**.** Given a utility function $u : Y \to \mathbb{R}$ and probability measures $\mu, \nu \in \Delta(\mathcal{Y})$, the *expected utility* of $\mu$ is $\mathbb{E}_\mu[u]$.

$\mu$ is *preferred* to $\nu$ if $\mathbb{E}_\mu[u] \geq \mathbb{E}_\nu[u]$, and *strictly preferred* if $\mathbb{E}_\mu[u] > \mathbb{E}_\nu[u]$.

**Definition 4.2.6** (Decision rule)**.** Given a see-to map $\mathbb{T} : \mathsf{H} \times D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$, a *decision rule* is a Markov kernel $X \to \Delta(\mathcal{D})$. A *deterministic decision rule* is a decision rule that is deterministic.

Define deterministic Markov kernels

Expected utility together with a decision rule gives rise to the definition of *risk*, which connects CSDT to classical statistical decision theory (SDT). For historical reasons, risks are minimised while utilities are maximised.

**Definition 4.2.7** (Risk)**.** Given a see-to map $\mathbb{T} : \mathsf{H} \times D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$, a utility $u : Y \to \mathbb{R}$ and the set of decision rules $\mathcal{U}$, the *risk* is a function $l : \mathsf{H} \times \mathcal{U} \to \mathbb{R}$ given by

$$R(\mathsf{H}, \mathbb{U}) := -\int_X \mathbb{U}_x \mathbb{T}^{\mathsf{Y}|\mathsf{DXH}}_{\cdot,x,\mathsf{H}} u d\mathbb{T}^{\mathsf{X}|\mathsf{H}}_{\mathsf{H}}(x) \tag{4.9}$$

for $\mathsf{H} \in \mathsf{H}$, $\mathbb{U} \in \mathcal{U}$. Here $\mathbb{U}_x \mathbb{T}^{\mathsf{Y}|\mathsf{DXH}}_{\cdot,x,\mathsf{H}} u$ is the product of the measure $\mathbb{U}_x$, the kernel $\mathbb{T}^{\mathsf{Y}|\mathsf{DXH}}_{\cdot,x,\mathsf{H}} : D \to \Delta(\mathcal{Y})$ and the function $u$.

The loss induces a partial order on decision rules. If for all $\mathsf{H}$, $l(\mathsf{H}, \mathbb{U}) \leq l(\mathsf{H}, \mathbb{U}')$ then $\mathbb{U}$ is at least as good as $\mathbb{U}'$. If, furthermore, there is some $\mathsf{H}_0$ such that $l(\mathsf{H}_0, \mathbb{U}) < l(\mathsf{H}_0, \mathbb{U}')$ then $\mathbb{U}$ is preferred to $\mathbb{U}'$.

**Definition 4.2.8** (Induced statistical decision problem)**.** A see-do model $\mathbb{T} : \mathsf{H} \times D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ along with a utility $u$ induces the *statistical decision problem* $(\mathsf{H}, \mathcal{U}, R)$ with states $\mathsf{H}$, decisions $\mathcal{U}$ and risks $R$.

Statistical decision problems usually define the risk via the loss, but it is only possible to define a loss with a decomposable model. We don't actually need a loss, though: the complete class theorem still holds via the induced risk and Bayes risk

We develop causal statistical decision problems (CSDPs) inspired by statistical decision problems (SDPs) of Wald (1950). CSDPs differ from SDPs in that our preferences (i.e. utility or loss) are known less directly in former case. We show that every SDP can be represented by a CSDP and that the converse is sometimes but not always possible. We show that an analogoue of the fundamental *complete class theorem* of SDPs applies to the class of CSDPs that can be represented by SDPs, but whether such a theorem applies more generally is an open question.

Following (Ferguson, 1967), we consider SDPs and CSDPs to represent normal form two person games. At the most abstract level the games represent the options and possible payoffs available to the decision maker, and this representation allows us to compare the two types of problem. In their more detailed versions, CSDPs and SDPs differ in their representation of the state of the world and in the type of function that represents preferences. These differences are summarised in Table **??**.

**Definition 4.2.9** (Normal form two person game)**.** A normal form game is a triple $\langle \mathcal{S}, A, L \rangle$ where $\mathcal{S}$ and $A$ are arbitrary sets and $L : \mathcal{S} \times A \to [0, \infty)$ is a loss function.

The set $\mathcal{S}$ is a set of possible states that the environment may occupy and $A$ is a set of actions the decision maker may take. The decision maker seeks an action in $A$ that minimises the loss $L$. Generally there is no action that minimises the loss for all environment states. A minimax solution is an action that minimises the worst case loss: $a^*_{mm} = \arg\min_{a \in A} [\sup_{s \in \mathcal{S}} L(s, a)]$.

If the set $\mathcal{S}$ is equipped with a $\sigma$-algebra $\mathcal{S}$ and a probability measure $\xi \in \Delta(\mathcal{S})$ which we will call a "prior", a Bayes solution minimizes the expected risk with respect to $\xi$: $a^*_{ba} = \arg\min_{a \in A} \int_{\mathcal{S}} L(s, a) \xi(ds)$.

**Definition 4.2.10** (Admissible Action)**.** Given a normal form two person game $\langle \mathcal{S}, A, L \rangle$, an action $a \in A$ is *strictly better* than $a' \in A$ iff $L(s, a) \leq L(s, a')$ for all $s \in \mathcal{S}$ and $L(s_0, a) < L(s_0, a')$ for some $s_0 \in \mathcal{S}$. If only the first holds, then $a$ is as good as $a'$. An *admissible action* is an action $a \in A$ such that there is no action strictly better than $A$.

**Definition 4.2.11** (Complete Class)**.** A class $C$ of decisions is a *complete class* if for every $a \notin C$ there is some $a' \in C$ that is strictly better than $a$.

$C$ is an *essentially complete* class if for every $a \notin C$ there is some $a' \in C$ that is as good as $a$.

A statistical decision problem represents a normal form two-person game where the available actions are *decision functions* that output a decision given data, the states of the environment are associated with probability measures on some measurable space and we assume a loss expressing preferences over decisions and states is known.

**Definition 4.2.12** (Statistical Experiment)**.** A *statistical experiment* relative to a set $\Theta$, a measurable space $(E, \mathcal{E})$ and a map $m : \Theta \to \Delta(\mathcal{E})$ is a multiset $\mathcal{H} = \{\mu_\theta | \theta \in \Theta\}$ where $\mu_\theta := m(\theta)$. The set $\Theta$ indexes the "state of nature".

**Definition 4.2.13** (Statistical Decision Problem)**.** A statistical decision problem (SDP) is a tuple $\langle \Theta, (\mathcal{H}, m), D, \ell \rangle$. $\mathcal{H} \subset \Delta(\mathcal{E})$ is a statistical experiment relative to states $\Theta$, space $(E, \mathcal{E})$ and map $m : \Theta \to \Delta(\mathcal{E})$, $D$ is the set of available decisions with some $\sigma$-algebra $\mathcal{D}$ and $\ell : \Theta \times D \to \mathbb{R}$ is a loss function where $\ell(\theta, \cdot)$ is measurable with respect to $\mathcal{D}$ and $\mathcal{B}(\mathbb{R})$.

Denote by $\mathcal{J}$ the set of stochastic decision functions $E \to \Delta(\mathcal{D})$. For $J \in \mathcal{J}$ and $\mu_\theta \in \mathcal{H}$, the risk $R : \Theta \times \mathcal{J} \to [0, \infty)$ is defined as $R(J, \theta) = \int_D \ell(\theta, y) \mu_\theta J(dy)$. The triple $\langle \Theta, \mathcal{J}, R \rangle$ forms a two player normal form game.

The loss function $\ell$ expresses preferences over general (state, decision) pairs. It may be the case that our preferences are most directly known over future states of the world - we know which results of our decisions are desirable and which are undesirable, which we represent with a *utility function*. In this case, if we are to induce preferences over the possible decisions, that we have a model that is more informative than a statistical experiment. In particular, we require each state of nature to be associated with both a distribution over the given information and a map from decisions to distributions over results - we call this map a *consequence*, and the object that pairs a distribution and a consequence with each state of the world a *causal theory*.

**Definition 4.2.14** (Consequences)**.** Given a measurable result space $(F, \mathcal{F})$ and a measurable decision space $(D, \mathcal{D})$, a Markov kernel $\kappa : D \to \Delta(\mathcal{F})$ is a *consequence mapping*, or just a *consequence*.

**Definition 4.2.15** (Causal state)**.** Given a consequence $\kappa : D \to \Delta(\mathcal{F})$, a measurable observation space $(E, \mathcal{E})$ and some distribution $\mu \in \Delta(\mathcal{E})$, the pair $(\kappa, \mu)$ is a *causal state* on $E, D$ and $F$. We refer to $\kappa$ as the consequence and $\mu$ as the observed distribution.

In many cases the observation space $E$ and the results space $F$ might coincide. However, these spaces are defined by different aspects of the given information: the former is fixed by what observations are available and the latter by which parts of the world are relevant to the investigator's preferences (see Theorems **??** and **??**), and there is not a clear reason to insist that these spaces should always be the same.

**Definition 4.2.16** (Causal Theory)**.** A causal theory $\mathcal{T}$ is a set of causal states sharing the same decision, observation and outcome spaces. We abuse notation to assign the "type signature" $\mathcal{T} : E \times D \twoheadrightarrow F$ for a causal theory with observed distributions in $\Delta(\mathcal{E})$ and consequences of type $D \to \Delta(\mathcal{F})$. The causal states of a theory $\mathcal{T}$ may be associated with a master set of states $\Theta$, but in contrast to a statistical experiment this is not necessary to define the basic associated decision problem.

**Definition 4.2.17** (Causal Statistical Decision Problem)**.** A causal statistical decision problem (CSDP) is a triple $\langle \mathcal{T}, D, u \rangle$. $\mathcal{T}$ is a causal theory on $D \times E \twoheadrightarrow F$, $D$ is the decision set with $\sigma$-algebra $\mathcal{D}$ and $u : F \to \mathbb{R}$ is a measurable utility function expressing preference over the results of decisions.

Define the canonical loss $L : \mathcal{T} \times D \to \mathbb{R}$ by $L : (\kappa, \mu), y \mapsto -\mathbb{E}_{\gamma\kappa}[u]$. This change conforms with the conventions that utilities are maximised while losses are minimised.

Given a decision function $J \in \mathcal{J}$ and $(\kappa, \mu) \in \mathcal{T}$, we define the risk $R : \mathcal{T} \times \mathcal{J} \to [0, \infty)$ by $R(\kappa, \mu, J) := L((\kappa, \mu), \mu J)$. The triple $\langle \mathcal{T}, \mathcal{J}, R \rangle$ is a normal form two person game.

The loss and the utility differ in that the loss expresses per-state preferences while the utility expresses state independent preferences. While we choose the loss to be a particular function of the utility here, it is possible to allow losses to be a more general class of functions of the utility and state without altering the preference ordering of a CSDP under minimax or Bayes decision rules. Given arbitrary $f : \mathcal{T} \to \mathbb{R}$, define $l : \mathcal{T} \times D \to \mathbb{R}$ by $l : (\kappa, \mu, y) \mapsto af(\kappa, \mu) + b\mathbb{E}_{\delta_y \kappa}[u]$. We can define a loss (relative to $f$) $L : \mathcal{T} \times \Delta(\mathcal{D}) \to [0, \infty]$ by

$$L((\kappa, \mu), \gamma) := \mathbb{E}_\gamma[l(\kappa, \mu, \cdot)] \tag{4.10}$$

$$= af(\kappa, \mu) - b\mathbb{E}_{\gamma\kappa}[u] \tag{4.11}$$

$$\tag{4.12}$$

For $(\kappa, \mu) \in \mathcal{T}$, $\gamma \in \Delta(\mathcal{D})$ and $a \in \mathbb{R}$, $b \in \mathbb{R}^+$.

A common example of a loss of the type above is the *regret*, which takes $a = b = 1$ and $f(\kappa, \mu) = \sup_{\gamma' \in \Delta(\mathcal{D})} \mathbb{E}_{\gamma'\kappa}[u]$. Because expected utility preserves preference orderings under positive affine transformations, the ordering of preferences given a particular state is not affected by the choices of $a, b$ and $f$, nor is the Bayes ordering of preferences given some prior $\xi$ over $\mathcal{T}$. While it may be possible to formulate decision rules for which the choices of $a, b$ and $f$ do matter, we will take these properties as sufficient to allow us to choose $a = 0$ and $b = 1$. More general classes of loss are of interest. *Regret theory*, for example, is a straightforward generalisation of the losses discussed here and is a prominent alternative to expected utility theory (Loomes and Sugden, 1982).

There are obvious similarities between SDPs and CSDPs: both have the same high level representation as a two person game which is arrived at by taking the expectation of a loss with respect to a decision function. In fact, if we consider two decision problems to be the same if they have the same representation as a two player game, we find that CSDPs are a special case of SDPs.

**Theorem 4.2.18** (CSDPs are a special case of SDPs)**.** *Given any CSDP $\alpha = \langle \mathcal{T}, D, u \rangle$ with two player game representation $\langle \mathcal{T}, \mathcal{J}, R \rangle$, there exists an SDP $\langle \mathcal{T}, (\mathcal{H}, m), D, \ell \rangle$ with the same representation as a two player game.*

*Proof.* Let $m : \mathcal{T} \to \mathcal{H}$ be defined such that $m : (\kappa, \mu) \mapsto \mu$ for $(\kappa, \mu) \in \mathcal{H}$. Define $\ell : \mathcal{T} \times D \to \mathbb{R}$ by $\ell : ((\kappa, \mu), y) \mapsto -\mathbb{E}_{\delta_y \kappa}[u]$. Let $R'((\kappa, \mu), J) = \mathbb{E}_{\mu J}[\ell(\theta, \cdot)]$.

Then

$$R'((\kappa,\mu),J) = -\int_D \mathbb{E}_{\delta_y \kappa}[u]\mu J(dy) \tag{4.13}$$

$$= -\int_D \int_F u(x)\kappa(y;dx)\mu J(dy) \tag{4.14}$$

$$= -\int_F u(x)\mu J\kappa(dx) \tag{4.15}$$

$$= R((\kappa,\mu),J) \tag{4.16}$$

□

The converse is not true, as the set $\Theta$ in an SDP is of an arbitrary type and may not be a causal theory. However, it is possible for any SDP with environmental states $\Theta$ to find a CSDP with causal theory $\mathcal{T}$ such that the games represented by each decision problem are related by a surjective map $f : \Theta \to \mathcal{T}$ which associates each state of nature with a causal state. We call such a map a *reduction* from an SDP to a CSDP.

**Definition 4.2.19** (Reduction)**.** Given normal form two person games $\alpha = \langle \mathcal{S}^\alpha, A, L^\alpha \rangle$ and $\beta = \langle \mathcal{S}^\beta, A, L^\beta \rangle$, $f : \mathcal{S}^\alpha \to \mathcal{S}^\beta$ is a *reduction* from $\alpha$ to $\beta$ if, defining the image $f(\mathcal{S}^\alpha) = \{f(\theta) | \theta \in \mathcal{S}^\alpha\}$, we have $\langle \mathcal{S}^\beta, A, L^\beta \rangle = \langle f(\mathcal{S}^\alpha), A, L^\alpha \circ (f \otimes I_A) \rangle$.

**Theorem 4.2.20** (SDP can be reduced to a CSDP)**.** *Given any SDP $\langle \Theta, (\mathcal{H}, m), D, \ell \rangle$ represented as the game $\alpha = \langle \Theta, \mathcal{J}, R \rangle$, there exists a CSDP $\langle \mathcal{T}, D, u \rangle$ represented as the game $\beta = \langle \mathcal{T}, \mathcal{J}, R' \rangle$ such that there is some reduction $f : \Theta \to \mathcal{T}$ from $\alpha$ to $\beta$.*

*Proof.* Take $\mathcal{H} \subset \Delta(\mathcal{E})$ and define $f : \Theta \to \Delta(\mathcal{E}) \times \Delta(\mathcal{B}(\mathbb{R}))^D$ by $f : \theta \mapsto (y \mapsto \delta_{l(\theta,y)}, \mu_\theta)$. Noting that $y \mapsto \delta_{l(\theta,y)}$ is a Markov kernel $D \to \Delta(\mathcal{B}(\mathbb{R}))$, the image $f(\Theta)$ is a causal theory $E \times D \twoheadrightarrow \mathbb{R}$. Consider the CSDP $\langle f(\Theta), D, -I_{(\mathbb{R})} \rangle$. Then, letting $R'$ denote the risk associated with this theory

$$R'((\kappa,\mu),J) = -\int_{\mathbb{R}} \int_D (-x)\delta_{l(\theta,y)}(dx)\mu_\theta J(dy) \tag{4.17}$$

$$= \int_D l(\theta,y)\mu_\theta J(dy) \tag{4.18}$$

$$= R(\Theta, J) \tag{4.19}$$

□

The fundamental *complete class theorem* of SDPs establishes that there are no decision rules that dominate the set of all Bayes rules under some regularity assumptions. By theorem 4.2.18, this must also be true of CSDPs.

**Theorem 4.2.21** (Complete class theorem (CSDP))**.** *Given any CSDP $\alpha := \langle \mathcal{T}, D, u \rangle$ with two player game representation $\langle \mathcal{T}, \mathcal{J}, R \rangle$, if $|\mathcal{T}| < \infty$ and $\inf_{J \in \mathcal{J}, (\kappa,\mu) \in \mathcal{H}} R((\kappa,\mu),J) >$*

*$-\infty$, then the set of all Bayes decision functions is a complete class for $\alpha$ and the set of all admissible Bayes decision functions is a minimal complete class for $\alpha$.*

*Proof.* By theorem 4.2.18, there exists an SDP $\beta$ such that $\alpha$ and $\beta$ have the same representation as a two player game. By assumption, $\beta$ has a finite set of states and a risk function that is bounded below. Therefore the Bayes rules on $\alpha$ are a complete class and admissible Bayes rules are a minimal complete class for the problem $\langle \mathcal{T}, \mathcal{J}, R \rangle$ (Ferguson, 1967). □

# Chapter 5

# See-do models, interventions and counterfactuals

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

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

## 5.2 Interpretations of the choice set

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

## 5.3 Causal Bayesian Networks as see-do models

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

- Example: sets of CBNs and d-separation

## 5.4 Unit Potential Outcomes models

- Counterfactual random variables 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 $\mathsf{H}_1$: the switch on the wall controls my light, $\mathsf{H}_2$: the switch on the wall does not control my light. Then, given $\mathsf{H}_1$ 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 $\mathsf{H}_2$, 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 $\mathsf{H}$ 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).

### 5.4.1 D-causation

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

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

Section **??** explores the difficulty of defining "objective causation" without reference to a set of choices. $D$ need not be interpreted as the set of choices available to an agent, but however we want to interpret it, all existing examples of causal models seem to require this set.

See Section 2.0.4 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 5.4.1** ($D$-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), $\mathsf{Y}_1$ $D$-causes $\mathsf{Y}_2$ iff $\mathsf{Y}_2 \perp\!\!\!\perp_{\mathbb{C}} \mathsf{D}|\mathsf{Y}_1\mathsf{H}$.

### 5.4.2 D-causation vs Limited Unresponsiveness

Heckerman and Shachter study deterministic "consequence models". Furthermore, what we call hypotheses $h \in H$, Heckerman and Schachter call states $s \in S$. Heckerman and Shachter's notion of causation is defined by *limited unresponsiveness* rather than *conditional independence*, which depends on a partition of states rather than a particular hypothesis.

**Definition 5.4.2** (Limited unresponsiveness)**.** Given states $S$, deterministic consequence models $\mathbb{C}_s : D \to \Delta(F)$ for each $s \in A$ and a random variables $\mathsf{Y}_1 : F \to Y_1$, $\mathsf{Y}_2 : F \to Y_2$, $\mathsf{Y}_1$ is unresponsive to $\mathsf{D}$ in states limited by $\mathsf{Y}_2$ 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 $\mathsf{Y}_1 \not\leftarrow_{\mathsf{Y}_2} \mathsf{D}$

**Lemma 5.4.3** (Limited unresponsiveness implies $D$-causation)**.** *For deterministic consequence models,* $\mathsf{Y}_1 \not\leftarrow_{\mathsf{Y}_2} \mathsf{D}$ *implies* $\mathsf{Y}_2$ *$D$-causes* $\mathsf{Y}_1$.

*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 $A \in \mathcal{Y}_2$, $B \in \mathcal{Y}_1$, $s \in S$, $d \in D$:

$$\mathbb{C}^{\mathsf{Y}_2|\mathsf{SD}}_{(s,d)} \curlyvee (\mathbb{M} \otimes \mathrm{Id}) = \int_A \mathbb{M}(y_2', d, s; B) d\delta_{y_2(s,d)}(y_2') \tag{5.1}$$

$$= \int_A \delta_{f(y_2',s)}(B) d\delta_{y_2(s,d)}(y_2') \tag{5.2}$$

$$= \delta_{f(y_2(s,d),s)}(B)\delta_{y_2(s,d)}(A) \tag{5.3}$$

$$= \delta_{y_1(s,d)}(B)\delta_{y_2(s,d)}(A) \tag{5.4}$$

$$= \delta_{y_2(s,d)} \otimes \delta_{y_1(s,d)}(A \times B) \tag{5.5}$$

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

define this

However, despite limited unresponsiveness implying $D$-causation, it does not imply $D$-causation in mixtures of states. Suppose $D = \{0, 1\}$ where 1 stands for "toggle light switch" and 0 stands for "do nothing". Suppose $S = \{[0,0], [0,1], [1,0], [1,1]\}$ where $[0,0]$ represents "switch initially off, mains off" the other states generalise this in the obvious way. Finally, $\mathsf{F} \in \{0, 1\}$ is the final position of the switch and $\mathsf{L} \in \{0, 1\}$ is the final state of the light. We have

$$\mathbb{C}^{\mathsf{LF}|\mathsf{DS}}_{d,[i,m]} = \delta_{(d \text{ XOR } i) \text{ AND } m} \otimes \delta_{(d \text{ XOR } i) \text{ AND } m} \tag{5.6}$$

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:

$$\mathbb{C}_\gamma = \frac{1}{4}\mathbb{C}_{\cdot,[0,0]} + \frac{1}{4}\mathbb{C}_{\cdot,[0,1]} + \frac{1}{2}\mathbb{C}_{\cdot,[1,1]} \tag{5.7}$$

$$\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma = \frac{1}{4}\begin{bmatrix}1&0\\0&1\end{bmatrix} \otimes \begin{bmatrix}1&0\\1&0\end{bmatrix} + \frac{1}{4}\begin{bmatrix}1&0\\0&1\end{bmatrix} \otimes \begin{bmatrix}1&0\\0&1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}0&1\\1&0\end{bmatrix} \otimes \begin{bmatrix}0&1\\1&0\end{bmatrix} \tag{5.8}$$

Then

$$[1,0]\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma = \frac{1}{4}[0,1] \otimes [1,0] + \frac{1}{4}[0,1] \otimes [0,1] + \frac{1}{2}[1,0] \otimes [1,0] \tag{5.9}$$

$$[1,0]\curlyvee(\mathbb{C}^{\mathsf{F}|\mathsf{D}}_\gamma \otimes \mathbb{C}^{\mathsf{L}|\mathsf{D}}_\gamma) = (\frac{1}{2}[0,1] + \frac{1}{2}[1,0]) \otimes (\frac{1}{4}[0,1] + \frac{3}{4}[1,0]) \tag{5.10}$$

$$\implies [1,0]\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma \neq [1,0]\curlyvee(\mathbb{C}^{\mathsf{F}|\mathsf{D}}_\gamma \otimes \mathbb{C}^{\mathsf{L}|\mathsf{D}}_\gamma) \tag{5.11}$$

Thus under the prior $\gamma$, $\mathsf{F}$ does not $D$-cause $\mathsf{L}$ even though $\mathsf{F}$ $D$-causes $\mathsf{L}$ 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 $\mathsf{X}$ $D$-causes $\mathsf{Y}$ in every $\mathsf{H} \in S$, $\mathsf{X}$ does not necessarily $D$-cause $\mathsf{Y}$ in mixtures of states in $S$. For this reason, we do not say that $\mathsf{X}$ $D$-causes $\mathsf{Y}$ in $S$ if $\mathsf{X}$ $D$-causes $\mathsf{Y}$ in every $\mathsf{H} \in S$, and in this way we differ substantially from Heckerman and Shachter (1995).

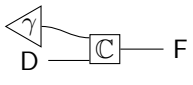
define this

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 $\mathsf{X}$ $D$-causes $\mathsf{Y}$ relative to $\gamma$ iff $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{C}_\gamma} \mathsf{D}|\mathsf{X}$.

Theorem 5.4.4 shows that under some conditions, $D$-causation can hold for arbitrary mixtures over subsets of the hypothesis class $\mathsf{H}$.

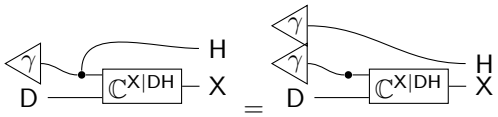
**Theorem 5.4.4** (Universal $D$-causation)**.** *If* $\mathsf{X} \perp\!\!\!\perp \mathsf{H}|\mathsf{D}$ *for all* $\mathsf{H}, \mathsf{H}' \in S \subset \mathsf{H}$ *and* $\mathsf{X}$ *$D$-causes* $\mathsf{Y}$ *in all* $\mathsf{H} \in S$*, then* $\mathsf{X}$ *$D$-causes* $\mathsf{Y}$ *with respect 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

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

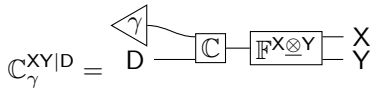
(5.12)

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

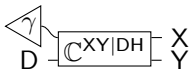



(5.13)

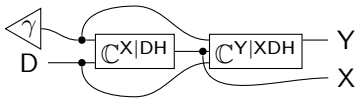
Also



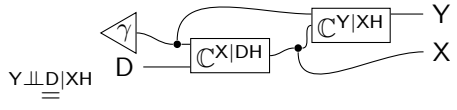

(5.14)



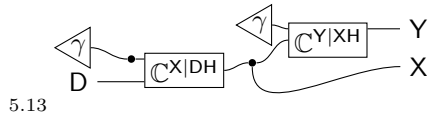

(5.15)



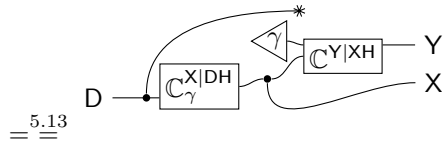

(5.16)




(5.17)




(5.18)




(5.19)

Equation 5.19 establishes that $(\gamma \otimes \mathbf{Id}_X \otimes \text{*}_D)\mathbb{C}^{Y|XH}$ is a version of $\mathbb{C}^{Y|XD}_{\gamma}$, and thus $Y \perp\!\!\!\perp_{\mathbb{C}_\gamma} D|X$.

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

$$H \perp\!\!\!\perp D \wedge H \perp\!\!\!\perp X|D \implies H \perp\!\!\!\perp XD \tag{5.20}$$
$$\implies H \perp\!\!\!\perp D|X \tag{5.21}$$
$$D \perp\!\!\!\perp H|X \wedge D \perp\!\!\!\perp Y|XH \implies D \perp\!\!\!\perp Y|X \tag{5.22}$$
$$\implies Y \perp\!\!\!\perp D|X \tag{5.23}$$

□

### 5.4.3 Properties of D-causation

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

$$\mathbb{C}_{\mathsf{H}}^{\mathsf{X}|\mathsf{D}} = \quad \mathsf{D} \;\boxed{\mathbb{C}^{\mathsf{X}|\mathsf{D}}}\!\!-\!\!\boxed{\mathbb{C}^{\mathsf{Y}|\mathsf{X}}}\; \mathsf{Y} \tag{5.24}$$

This follows from version (2) of Definition 2.0.35:

$$\mathbb{C}_{\mathsf{H}}^{\mathsf{X}|\mathsf{D}} = \quad \mathsf{D} \;\boxed{\mathbb{C}^{\mathsf{X}|\mathsf{D}}}\!\!-\!\!\boxed{\mathbb{C}^{\mathsf{Y}|\mathsf{XD}}}\; \mathsf{Y} \tag{5.25}$$

$$= \quad \mathsf{D} \;\boxed{\mathbb{C}^{\mathsf{X}|\mathsf{D}}}\!\!-\!\!\boxed{\mathbb{C}^{\mathsf{Y}|\mathsf{X}}}\; \mathsf{Y} \tag{5.26}$$

$$= \quad \mathsf{D} \;\boxed{\mathbb{C}^{\mathsf{X}|\mathsf{D}}}\!\!-\!\!\boxed{\mathbb{C}^{\mathsf{Y}|\mathsf{X}}}\; \mathsf{Y} \tag{5.27}$$

D-causation is not transitive: if $\mathsf{X}$ D-causes $\mathsf{Y}$ and $\mathsf{Y}$ D-causes $\mathsf{Z}$ then $\mathsf{X}$ doesn't necessarily D-cause $\mathsf{Z}$.

Pearl's "front door adjustment" and general identification results make use of composing "sub-consequence-kernels" like this. Show, if possible, that Pearl's "sub-consequence-kernels" obey *D*-causation like relations

Does this "weak D-causation" respect mixing under the same conditions as regular D-causation?

### 5.4.4 Decision sequences and parallel decisions

Just as observations $\mathsf{X}$ can be a sequence of random variables $\mathsf{X}_1$, $\mathsf{X}_2$, ..., $\mathsf{D}$ can be a sequence of "sub-choices" $\mathsf{D}_1$, $\mathsf{D}_2$, ... . Note that by positing such a sequence there is no requirement that $\mathsf{D}_1$ comes "before" $\mathsf{D}_2$ in any particular sense.

## 5.5 Existence of counterfactuals

I'm struggling with how to explain this well.

"Counterfactual" or "potential outcomes" models in the causal inference literature are consequence models where choices can be considered in *parallel*.

Before defining parallel choices, we will consider a "counterfactual model" without parallel choices. Consider the following definitions, first from Pearl (2009) pg. 203-204. I have preserved his notation, including not using any special fonts for things called "variables" because this term is used interchangeably with "sets of variables" and using special fonts for variables might give the impression that these should be treated as different things while using special fonts for sets of variables is inconsistent with my usual notation.

The real solution here is that Pearl's "variable sets" are actually "coupled variables", see Definition 2.0.9, 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 $Y_{\cdot}(\cdot) : D \times U \to Y$ (distinct from $Y$, which appears to be a function $U \to$ something) and calls $Y_{\cdot}(\cdot)$ 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" $\mathbf{g} := \mathsf{V}_{\cdot}(\cdot)$, 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.

# Chapter 6

# Imitablity and inferring causes from data

## 6.1 Assumptions enabling learning

- Recall chapter 4: with no assumptions, no learning possible
- What kind of assumptions can be made? Chapter 5: setting, d-separation and ignorability
- Here we investigate:
  - Conditional setting (chapter 4: decomposability)
  - Imitability (chapter 3: double exchangeability)

## 6.2 Imitability

- Define "perfect knowledge" imitability
- Alternatively, could assume "approximate imitability given actual knowledge" (but we don't)
  - Approximate imitability doesn't require double exchangeability
- Define conditional setting

## 6.3 Identification with imitability

- Define "pre-intervention distribution RV"
- Define $D$-control
- Define identification

  - Explain weak conditional setting, need for alternative notion of identification

- Assumption: pre-intervention + ordinary RV $D$-controls consequence of interest; thm follows
- Reduction to SWIGs and hard intervention CBNs via actuator randomisation
  - SWIG "counterfactual" RVs are different to regular counterfactuals and can be defined also in "counterfactual-free" CBNs
- $D$-control can be deduced from conditional independence under imitability + assumptions of coverage
- Similar to Peters et al. (2016)
- What about faithfulness?? I've no idea

# Chapter 7

# Causal relationships on God's computer

## 7.1 Are we trying to understand consequences or actions or objective causal relationships?

- Whatever a cause is, if Y is by definition f(X,Z) then it seems very reasonable to call X and Z causes of Y (example: body mass index)
- Indeed this is one of the "intuitive justifications" of SEM
- But also, we can't do any interventions on Y (theorem)
- But most random variables are by definition functions of other variables (e.g. height in m: time in seconds it takes for light to travel from a person's feet to their head x SI speed of light)
- Also all random variables are by definition f(Omega)
- Therefore one can only intervene on the whole universe?
- There might be "objective causal relationships", but intervention doesn't seem like a strong candidate for defining them
- In contrast, there's no apparent contradiction in assuming:
  - We can control our BMI to some extent by adopting diets, exercise
  - If we survey data from an appropriately designed experiment, we can perhaps imitate the results and our health outcomes are perhaps D-controlled by (actual BMI, business as usual BMI)
  - Probably all of our options only affect weight and not height, but if the above holds we don't need to worry about this, nor about different effects on muscle mass, fat mass, other mass

Disorganised cut and paste follows

### 7.1.1 Necessary relationships

The relationship between a person's body mass index, their weight and their height defines what body mass index is. A fundamental claim of ours is that any causal model that defines "the causal effect of body mass index" should do so without reference to any submodel that violates this definitional relationship violation of the definition. This is an important assumption, and it rests on a judgement of what causal models ought to do. I think it is quite clear that when anyone asks for a causal effect, they expect that any operations required to define the causal effect *do not change the definitions of the variables they are employing.* While theories of causality have a role in sharpening our understanding of the term *causal effect*, the thing called a "causal effect" in an SCM should still respect some of our pre-theoretic intuitions about what causal effects are or else it should be called something else. "Causal effects" that depend on redefining variables do not respect pre-theoretic intuitions about what causal effects are:

- If I ask for the "causal effect of a person's BMI", I do not imagine that I am asking what would happen if someone's BMI were defined to be something other than their weight divided by their height
- If I ask for the "causal effect of a person's weight", I do not imagine that I am asking what would happen if someone's weight were not equal to their volume multiplied by their density
- If I ask for the "causal effect of a person's weight", I also do not imagine that I am asking what would happen if their weight were not equal to the weight of fat in their body plus the weight of all non-fat parts of their body
- If I ask for the "causal effect of taking a medicine", I do not imagine that I am asking what would happen if a person were declared to have taken a medicine independently of whatever substances have actually entered their body and how they entered

We will call relationships that have to hold *necessary relationships.* We provide the example of relationships that have to hold by definition as examples, but definitions may not be the only variety of necessary relationships. For example, one might also wish to stipulate that certain laws of physics are required to hold in all submodels.

If a causal model contains variables that are necessarily related, then an intervention on one of them must always change another variable in the relationship. If I change a person's weight, their height or BMI must change (or both). If I change their height, their weight or BMI must change and if I change their BMI then their weight or height must change. This conflicts with the usual acyclic definition of causal models, where the proposition that $\mathsf{A}$ causes $\mathsf{B}$ rules out the possibility that $\mathsf{B}$ or any of its descendents are a cause of $\mathsf{A}$. Thus in an

acyclic model it isn't possible for for an intervention on BMI to change weight or height and interventions on weight and height to also change BMI. Theroem 7.1.10 formalises this conflict for recursive structural causal models: for any set of variables that are necessarily related by a cyclic relationship, at least one of them has no hard interventions defined.

### 7.1.2 Recursive Structural Causal Models

We begin by showing that necessary relationships are incompatible with structural causal models.

**Definition 7.1.1** (Recursive Structural Causal Model)**.** A recursive structural causal model (SCM) is a tuple

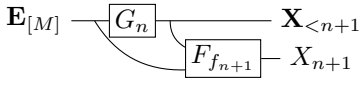
$$\mathcal{M} := \langle N, M, \mathbf{X}_{[N]}, \mathbf{E}_{[M]}, \{f_i | i \in [N]\}, \mathbb{P}_{\mathcal{E}} \rangle \tag{7.1}$$

where

- $N \in \mathbb{N}$ is the number of *endogenous variables* in the model
- $M \in \mathbb{N}$ is the number of *exogenous variables* in the model
- $\mathbf{X}_{[N]} := \{X_i | i \in [N]\}$ where, for each $i \in [N]$, $(X_i, \mathcal{X}_i)$ is a standard measurable space taking and the codomain of the $i$-th endogenous variable
- $\mathbf{E}_{[M]} := \{E_j | j \in [M]\}$ where, for $j \in [M]$, $E_j$ is a standard measurable space and the codomain of the $j$-th exogenous variable
- $f_i : \mathbf{X}_{<i} \times \mathbf{E}_{\mathcal{J}} \to X_i$ is a measurable function which we call *the causal mechanism controlling the $i$-th endogenous variable*
- $\mathbb{P}_{\mathcal{E}} \in \Delta(\mathbf{E}_{\mathcal{J}})$ is a probability measure on the space of exogenous variables

**Definition 7.1.2** (Observable kernel)**.** Given an SCM $\mathcal{M}$ with causal mechanisms $\{f_i | i \in [N]\}$, define the *observable kernel* $G_i : E \to \Delta(\mathbf{X}_{[i]})$ recursively:

$$G_1 = \tag{7.2}$$




$$G_{n+1} = \tag{7.3}$$

**Definition 7.1.3** (Joint distribution on endogenous variables)**.** The *joint distribution on endogenous variables* defined by $\mathcal{M}$ is $\mathbb{P}_{\mathcal{M}} := \mathbb{P}_{\mathcal{E}} G_N$ (which is the regular kernel product, see Definition 2.0.2). For each $i \in [N]$ define the random variable $\mathsf{X}_i : \mathbf{X}_{[N]} \to X_i$ as the projection map $\pi_i : (x_1, ..., x_i, .., x_N) \mapsto x_i$. By Lemma 7.1.4, $\underset{i \in [N]}{\otimes} \mathsf{X}_i = \mathrm{Id}_{\mathbf{X}_{[N]}}$, and so $\mathbb{P}_{\mathcal{M}}$ is the joint distribution of the variables $\{\mathsf{X}_i | i \in [N]\}$.

I use the notation $\mathbb{P}_{\mathcal{M}}$ rather than $\mathbb{P}_{\mathsf{X}_{[N]}}$ to emphasize the dependence on the model $\mathcal{M}$.

**Lemma 7.1.4** (Coupled product of all random variables is the identity)**.** $\underline{\otimes}_{i\in[N]}\mathsf{X}_i = \mathrm{Id}_{\mathbf{X}_{[N]}}$

*Proof.* for any $\mathbf{X} \in \mathbf{X}_{[N]}$,

$$\underline{\otimes}_{i\in[N]}\mathsf{X}_i(\mathbf{X}) = (\pi_1(\mathbf{X}), ..., \pi_N(\mathbf{X})) \tag{7.4}$$
$$= (x_1, ..., x_n) \tag{7.5}$$
$$= \mathbf{X} \tag{7.6}$$

□

**Definition 7.1.5** (Hard Interventions)**.** Let $\mathcal{M}$ be the set of all $SCMs$ sharing the indices, spaces and measure $\langle N, M, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{[M]}, \mathbb{P}_{\mathcal{E}}\rangle$. Note that the causal mechanisms are not fixed.

Given an SCM $\mathcal{M} = \langle N, M, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{[M]}, \{f_i | i \in [N]\}, \mathbb{P}_{\mathcal{E}}\rangle$ and $\mathcal{S} \subset [N]$, a *hard intervention* on $\mathsf{X}_{\mathcal{S}}$ is a map $Do_{\mathcal{S}} : \mathbf{X}_{\mathcal{S}} \times \mathcal{M} \to \mathcal{M}$ such that for $\mathbf{a} \in \mathbf{X}_S$, $Do_{\mathcal{S}}(\mathbf{a}, \mathcal{M}) = \langle N, M, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{[M]}, \{f'_i | i \in [N]\}, \mathbb{P}_{\mathcal{E}}\rangle$ where

$$f'_i = f_i \qquad i \notin \mathcal{S} \tag{7.7}$$
$$f'_i = \pi_i(\mathbf{a}) \qquad i \in \mathcal{S} \tag{7.8}$$

To match standard notation, we will write $\mathcal{M}^{do(\mathsf{X}_{\mathcal{S}}=\mathbf{a})} := Do_{\mathcal{S}}(\mathbf{a}, \mathcal{M})$

### 7.1.3 Recursive Structural Causal Models with Necessary Relationships

Necessary relationships are extra constraints on the joint distribution on endogenous variables defined by an SCM. For example, given an SCM $\mathcal{M}$ if the variable $\mathsf{X}_1$ represents weight, $\mathsf{X}_2$ represents height and $\mathsf{X}_3$ represents BMI then we want to impose the constraint that

$$\mathsf{X}_3 = \frac{\mathsf{X}_1}{\mathsf{X}_2} \tag{7.9}$$

$\mathbb{P}_{\mathcal{M}}$-almost surely.

**Definition 7.1.6** (Constrained Recursive Structural Causal Model (CSCM))**.** A CSCM $\mathcal{M} := \langle N, M, \mathbf{X}_{[N]}, \mathbf{E}_{[M]}, \{f_i | i \in [N]\}, \{r_i | i \in [N]\}, \mathbb{P}_{\mathcal{E}}\rangle$ is an SCM along with a set of *constraints* $r_i : \mathbf{X}_{[N]} \to X_i$.

If $\mathsf{X}_i = r_i(\mathsf{X}_{[N]})$ $\mathbb{P}_{\mathcal{M}}$-almost surely then $\mathsf{M}$ is *valid*, otherwise it is *invalid*.

We can recover regular SCMs by imposing only trivial constraints:

**Lemma 7.1.7** (CSCM with trivial constraints is always valid)**.** *Let $\mathcal{M}$ be a CSCM with the trivial constraints $r_i = \pi_i$ for all $i \in [N]$. Then $\mathcal{M}$ is valid.*

*Proof.* By definition 7.1.6, we require $\mathsf{X}_i = \mathsf{X}_i$, $\mathbb{P}_{\mathcal{M}}$-almost surely. $\mathsf{X}_i(\mathbf{X}) = \mathsf{X}_i(\mathbf{X})$ for all $\mathbf{X} \in \mathbf{X}_{[N]}$ and $P_{\mathcal{M}}(\mathbf{X}_{[N]}) = 1$, therefore $\mathcal{M}$ is valid. □

Call a constraint $r_i$ *cyclic* if $\mathsf{X}_i = r_i(\mathsf{X}_{[N]})$ implies there exists an index set $O \subset [N]$, $O \ni i$, such that for each $j \in O$, $\mathbf{b} \in \mathbf{X}_{O\setminus\{j\}}$ there exists $a \in X_j$ such that

$$\mathsf{X}_{O\setminus\{j\}} = \mathbf{b} \tag{7.10}$$
$$\implies \mathsf{X}_j = a \tag{7.11}$$

BMI is an example of a cyclic constraint if we insist that weight and height are always greater than 0. If $\mathsf{X}_3 = \frac{\mathsf{X}_1}{\mathsf{X}_2}$ then we have:

$$[\mathsf{X}_1, \mathsf{X}_2] = [b_1, b_2] \tag{7.12}$$
$$\implies \mathsf{X}_3 = \frac{b_1}{b_2} \tag{7.13}$$
$$[\mathsf{X}_2, \mathsf{X}_3] = [b_2, b_3] \tag{7.14}$$
$$\implies \mathsf{X}_1 = b_2 b_3 \tag{7.15}$$
$$[\mathsf{X}_1, \mathsf{X}_3] = [b_1, b_3] \tag{7.16}$$
$$\implies \mathsf{X}_2 = \frac{b_1}{b_3} \tag{7.17}$$

The following is a generally useful lemma that should probably be in basic definitions of Markov kernel spaces

**Lemma 7.1.8** (Projection and selectors)**.** *Given an indexed product space $\mathbf{X} := \prod_{i\in\mathcal{I}} X_i$ with ordered finite index set $\mathcal{I} \ni i$, let $\pi_i : \mathbf{X} \to X_i$ be the projection of the $i$-indexed element of $\mathbf{X} \in \mathbf{X}$.*

*Let $F_{\pi_i} : \mathbf{X} \to \Delta(\mathcal{X}_i)$ be the Markov kernel associated with the function $\pi_i$, $F_{\pi_i} : \mathbf{X} \mapsto \delta_{\pi_i(\mathbf{X})}$. Given $O \subset \mathcal{I}$, define the selector $S_i^O$:*

$$S_i^O = \begin{cases} \mathrm{Id}_{X_i} & i \in O \\ {}^*_{X_i} & i \notin O \end{cases} \tag{7.18}$$

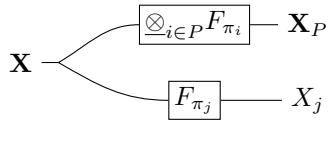
*Then $\underline{\otimes}_{i\in O} F_{\pi_i} = \otimes_{i\in\mathcal{I}} S_i^O$.*

*Proof.* Suppose $O$ is the empty set. Then the empty tensor product $\otimes_{i\in\emptyset} S_i$ and the empty coupled tensor product $\underline{\otimes}_{i\in\emptyset} F_{\pi_i}$ are both equal to ${}^*_{\mathbf{X}}$.

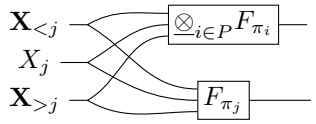
By definition of $F_{\pi_i}$, $F_{\pi_i} = \otimes_{i\in\mathcal{I}} S_i^{\{i\}}$.

Suppose for $P \subsetneq O$ with greatest element $k$ we have $\underline{\otimes}_{i\in P} F_{\pi_i} = \otimes_{i\in\mathcal{I}} S_i^P$, and suppose that $j$ is the next element of $O$ not in $P$.

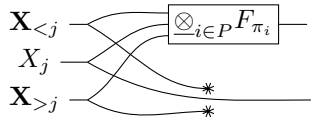
$$(\underline{\otimes}_{i\in P} F_{\pi_i}) \underline{\otimes} F_{\pi_j} = \tag{7.19}$$



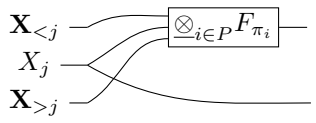

$$= \tag{7.20}$$




$$= \tag{7.21}$$

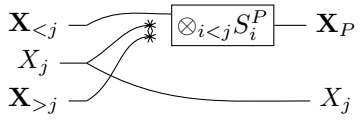



$$= \tag{7.22}$$

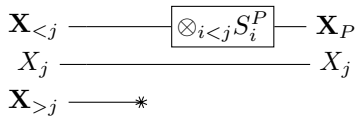



$$= \tag{7.23}$$

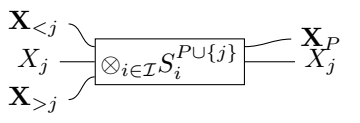
Because all elements of $P$ are less than $j$, the selector $S_k^P$ resolves to the discard map for $k > j$:




$$= \tag{7.24}$$




$$= \tag{7.25}$$




$$= \tag{7.26}$$

Where 7.26 follows from the definition of the selector $S_i^{P\cup\{j\}}$.
The proof follows by induction on the elements of $O$.

□

**Lemma 7.1.9** (Hard interventions do not affect the joint distributions of earlier variables)**.** *Given a CSCM* $\mathcal{M} = \langle N, M, \mathbf{X}_{[N]}, \mathbf{E}_{[M]}, \{f_i | i \in [N]\}, \{r_i | i \in$

$[N]\}, \mathbb{P}_{\mathcal{E}}\rangle$, *any* $k \in [N]$ *and any* $O \subset [k-1]$, $P_{\mathcal{M}}(\mathsf{X}_O) = P_{\mathcal{M}}^{do(\mathsf{X}_k)=a}(\mathsf{X}_O)$ *for all* $a \in X_k$.

*Proof.* Let $G_i^{\mathcal{Q}}$, $i \in [N]$ be the $i$-th iteration of the kernel defined in Equations 7.2 and 7.3 with respect to model $\mathcal{Q}$. Note that from Equation 7.3

$\mathbf{E}_{[M]}$ — $G_i^{\mathcal{Q}}$ — $\mathbf{X}_{<i}$, $*$

$$= \quad \mathbf{E}_{[M]} \text{ — } G_{i-1}^{\mathcal{Q}} \text{ — } \mathbf{X}_{<i}, \quad G_{i-1}^{\mathcal{Q}} \text{ — } F_{f_i} \text{ — } * \tag{7.27}$$

$$= G_{i-1}^{\mathcal{Q}} \tag{7.28}$$

It follows that

$$\mathbf{E}_{[M]} \text{ — } G_N \text{ — } \mathbf{X}_{<i}, * \quad = G_{i-1} \tag{7.29}$$

Because $f_i = f_i^{do(\mathsf{X}_k=a)}$ for $i < k$, we have

$$G_i^{\mathcal{M}} = G_i^{\mathcal{M}^{do(\mathsf{X}_k=a)}} \tag{7.30}$$

for all $i < k$. By lemma 7.1.8, for any $O \subset [k-1]$ we have $F_{\mathsf{X}_O} = \otimes_{i\in[N]} S_i^O$. As there are no elements of $O$ greater than or equal to $k$, the selector $S_i^O$ resolves to the discard for all $i >= k$. Thus $F_{\mathsf{X}_O} = (\otimes_{i\in[k-1]} S_i^O) \otimes *_{\mathbf{X}_{[N]\setminus[k-1]}}$. Defining $S_{[k-1]}^O := \otimes_{i\in[k-1]} S_i^O$, we have:

$$F_{\mathsf{X}_O} = \begin{array}{l} \mathbf{X}_{[k-1]} \text{ — } S_{[k-1]}^O \text{ — } \mathbf{X}_O \\ \mathbf{X}_{[N]\setminus[k-1]} \text{ — } * \end{array} \tag{7.31}$$

Thus

$$\mathbb{P}_{\mathcal{M}}(\mathsf{X}_O) = \mathbb{P}_{\mathcal{E}} G_N^{\mathcal{M}} F_{\mathsf{X}_O} \tag{7.32}$$

$\mathbb{P}_{\mathcal{E}}$ — $G_N^{\mathcal{M}}$ — $S_{[k-1]}^O$ — $\mathbf{X}_O$; $G_N^{\mathcal{M}}$ — $* (\mathbf{X}_{[N]\setminus[k-1]})$

$$\overset{7.31}{=} \tag{7.33}$$

$\mathbb{P}_{\mathcal{E}}$ — $G_{k-1}^{\mathcal{M}}$ — $S_{[k-1]}^O$ — $\mathbf{X}_O$

$$\overset{7.29}{=} \tag{7.34}$$

$\mathbb{P}_{\mathcal{E}}$ — $G_{k-1}^{\mathcal{M}^{do(\mathsf{X}_k=a)}}$ — $S_{[k-1]}^O$ — $\mathbf{X}_O$

$$\overset{7.30}{=} \tag{7.35}$$

$\mathbb{P}_{\mathcal{E}}$ — $G_N^{\mathcal{M}^{do(\mathsf{X}_k=a)}}$ — $S_{[k-1]}^O$ — $\mathbf{X}_O$; $G_N^{\mathcal{M}^{do(\mathsf{X}_k=a)}}$ — $* (\mathbf{X}_{[N]\setminus[k-1]})$

$$\overset{7.29}{=} \tag{7.36}$$

$$= P_{\mathcal{M}^{do(\mathsf{X}_k=a)}}(\mathsf{X}_O) \tag{7.37}$$

□

**Theorem 7.1.10** (Undefined hard interventions with cyclic constraints)**.** *Consier a CSCM* $\mathcal{M} = \langle N, M, \mathbf{X}_{[N]}, \mathbf{E}_{[M]}, \{f_i | i \in [N]\}, \{r_i | i \in [N]\}, \mathbb{P}_{\mathcal{E}} \rangle$ *with* $r_i$ *a cyclic constraint with respect to* $O \subset [N]$ *and the rest of the constraints trivial:* $r_j = \pi_j$, $j \neq i$, *and suppose* $\mathcal{M}$ *is valid.*

*If for each* $k \in O$, $\exists A \in \mathcal{X}_i$ *such that* $0 < \mathbb{P}_{\mathcal{M}}(\mathsf{X}_i \in A) < 1$ *then for at least one* $k \in O$ *all models given by a hard intervention on* $\mathsf{X}_k$ *are invalid.*

*Proof.* Choose $k$ to be the maximum element of $O$. By the assumption $\mathcal{M}$ is valid, we have $\mathsf{X}_i = r_i(\mathbf{X})$, $\mathbb{P}_{\mathcal{M}}$-almost surely. Let $B^A = \{\mathbf{b} \in \mathbf{X}_{O\setminus k} | \mathsf{X}_{O\setminus k} = \mathbf{b} \implies \mathsf{X}_k \in A\}$ and $B^{A^C} = \{\mathbf{b} \in \mathbf{X}_{O\setminus k} | \mathsf{X}_{O\setminus k} = \mathbf{b} \implies \mathsf{X}_k \notin A\}$.

$r_i$ holds on a set of measure 1, and wherever it holds $\mathsf{X}_{O\setminus\{k\}}$ is either in $B^A$ or $B^{A^C}$. Thus $\mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O\setminus\{k\}} \in B^A \cup B^{A^C}) = 1$.

$B^A$ and $B^{A^C}$ are disjoint.

By construction, $\mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O\setminus\{k\}} \in B^A \ \& \ \mathsf{X}_k \in A) = \mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O\setminus\{k\}} \in B^A)$ and $\mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O\setminus\{k\}} \in B^{A^C} \ \& \ \mathsf{X}_k \in A^C) = \mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O\setminus\{k\}} \in B^{A^C})$

By additivity, $\mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O\setminus\{k\}} \in B^A \ \& \ \mathsf{X}_k \in A) + \mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O\setminus\{k\}} \notin B^A \ \& \ \mathsf{X}_k \in A) = P_{\mathcal{M}}(\mathsf{X}_k \in A)$.

By additivity agian

$$\mathbb{P}_{\mathcal{M}}\left(\mathsf{X}_{O\setminus\{k\}} \notin B^A \ \& \ \mathsf{X}_k \in A\right) = \mathbb{P}_{\mathcal{M}}\left(\mathsf{X}_{O\setminus\{k\}} \in B^{A^C} \ \& \ \mathsf{X}_k \in A\right) \tag{7.38}$$

$$+ \mathbb{P}_{\mathcal{M}}\left(\mathsf{X}_{O\setminus\{k\}} \in (B^{A^C} \cup B^A)^C \ \& \ \mathsf{X}_k \in A\right) \tag{7.39}$$

$$<= 0 + P_{\mathcal{M}}\left(\mathsf{X}_{O\setminus\{k\}} \in (B^{A^C} \cup B^A)^C\right) \tag{7.40}$$

$$= 0 \tag{7.41}$$

Thus $\mathbb{P}_{\mathcal{M}}\left(\mathsf{X}_{O\setminus\{k\}} \in B^A \ \& \ \mathsf{X}_k \in A\right) = P_{\mathcal{M}}(\mathsf{X}_k \in A) = P_{\mathcal{M}}(\mathsf{X}_{O\setminus\{k\}} \in B^A)$ and by an analogous argument $\mathbb{P}_{\mathcal{M}}(\mathsf{X}_{O\setminus\{k\}} \in B^{A^C}) = P_{\mathcal{M}}(\mathsf{X}_k \in A^C)$.

Choose some $a \in A$, and consider the hard intervention $\mathcal{M}^{do(\mathsf{X}_k = a)}$. Suppose $\mathcal{M}^{do(\mathsf{X}_k = a)}$ is also valid. Then, as before, $\mathbb{P}_{\mathcal{M}^{do(\mathsf{X}_k = a)}}(\mathsf{X}_{O\setminus\{k\}} \in B^{A^C}) = \mathbb{P}_{\mathcal{M}^{do(\mathsf{X}_k = a)}}(\mathsf{X}_k \in A^C)$.

By definition of hard interventions, $f_k^{do(\mathsf{X}_k = a)} = a$. Thus $G_N^{\mathcal{M}^{do(\mathsf{X}_k = a)}} F_{\mathsf{X}_k}$ is the kernel $\mathbf{X} \mapsto \delta_a$ and it follows that $\mathbb{P}_{\mathcal{M}^{do(\mathsf{X}_k = a)}}(\mathsf{X}_k) = \delta_a$.

By lemma 7.1.9, $\mathbb{P}_{\mathcal{M}^{do(\mathsf{X}_k = a)}}(\mathsf{X}_{O\setminus\{k\}} \in B^{A^C}) = P_{\mathcal{M}}(\mathsf{X}_{O\setminus\{k\}} \in B^{A^C}) = P_{\mathcal{M}}(\mathsf{X}_k \in A^C) > 0$. But $P_{\mathcal{M}^{do(\mathsf{X}_k = a)}}(\mathsf{X}_k \in A^C) = \delta_z(\mathsf{X}_k \in A^C) = 0$, contradicting the assumption of validity of $\mathcal{M}^{do(\mathsf{X} = a)}$.

An analogous argument shows that all hard interventions $a' \in A^C$ are also invalid. □

### 7.1.4 Cyclic Structural Causal Models

It is not very surprising that acyclic causal models cannot accommodate cyclic constraints. Can cyclic causal models do so? While Bongers et al. (2016) has

develope a theory of cyclic causal models, cyclic are generally far less well understood than acyclic models. I show that the theory of cyclic models that Bongers has developed also fails to define hard interventions on variables subject to cyclic constraints. This does not rule out the possibility that there is some other way to define cyclic causal models that do handle these constraints, but I have not taken it upon myself to develop such a theory.

Haven't done any work from here on

We adopt the framework of cyclic structural causal models to make our arguments, adapted from Bongers et al. (2016). This is somewhat non-standard, but allows us to make a stronger argument for the impossibility of modelling arbitrary sets of variables using structural interventional models.

**Definition 7.1.11** (Structural Causal Model)**.** A structural causal model (SCM) is a tuple

$$\mathcal{M} := \langle \mathcal{I}, \mathcal{J}, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{\mathcal{J}}, \mathbf{f}_{\mathcal{I}}, \mathbb{P}_{\mathcal{E}}, \boldsymbol{E}_{\mathcal{J}} \rangle \tag{7.42}$$

where

- $\mathcal{I}$ is a finite index set of *endogenous variables*
- $\mathcal{J}$ is a finite index set of *exogenous variables*
- $\mathbf{X}_{\mathcal{I}} := \{X_i\}_{\mathcal{I}}$ where, for each $i \in \mathcal{I}$, $(X_i, \mathcal{X}_i)$ is a standard measurable space taking and the codomain of the $i$-th endogenous variable
- $\mathbf{E}_{\mathcal{J}} := \{E_j\}_{\mathcal{J}}$ where, for $j \in \mathcal{J}$, $E_j$ is a standard measurable space and the codomain of the $j$-th endogenous variable
- $\mathbf{f}_{\mathcal{I}} = \underline{\otimes}_{i\in\mathcal{I}} f_i$ is a measurable function, and $f_i : \mathbf{X}_{\mathcal{I}} \times \mathbf{E}_{\mathcal{J}} \to X_i$ is the causal mechanism controlling $\mathsf{X}_i$
- $\mathbb{P}_{\mathcal{E}} \in \Delta(\mathbf{E}_{\mathcal{J}})$ is a probability measure on the space of exogenous variables
- $\boldsymbol{E}_{\mathcal{J}} = \underline{\otimes}_{j\in\mathcal{J}} \mathsf{E}_j$ is the set of exogenous variables, with $\mathbb{P}_{\mathcal{E}} = \boldsymbol{E}_{\mathcal{J}\#} P_{\mathcal{E}}$ and $\mathsf{E}_j$ is the j-th exogenous variable with marginal distribution given by $\mathsf{E}_{j\#}\mathbb{P}_{\mathcal{E}}$

If for $\mathbb{P}_{\mathcal{E}}$-almost every $\mathbf{e} \in \mathbf{E}_{\mathcal{J}}$ there exists $\mathbf{X} \in \mathbf{X}_{\mathcal{I}}$ such that

$$\mathbf{X} = \mathbf{f}_{\mathcal{I}}(\mathbf{X}, \mathbf{e}) \tag{7.43}$$

Then an SCM $\mathcal{M}$ induces a unique probability space $(\mathbf{X}_{\mathcal{I}} \times \mathbf{E}_{\mathcal{J}}, \mathcal{X}_{\mathcal{I}} \otimes \mathcal{E}_{\mathcal{J}}, \mathbb{P}_{\mathcal{M}})$ (Bongers et al., 2016). If no such solution exists then we will say an SCM is invalid, as it imposes mutually incompatible constraints on the endogenous variables. It may be also the case that multiple solutions exist.

If an SCM induces a unique probability space then there exist random variables $\{\mathsf{X}_i\}_{i\in\mathcal{I}}$ such that, $P_{\mathcal{M}}$ almost surely Bongers et al. (2016):

$$\mathsf{X}_i = f_i(\boldsymbol{X}_{\mathcal{I}}, \boldsymbol{E}_{\mathcal{J}}) \tag{7.44}$$

Where $\boldsymbol{X}_{\mathcal{I}} = \underline{\otimes}_{i \in \mathcal{I}} \mathsf{X}_i$.

A structural causal model can be transformed by *mechanism surgery*. Given $\mathcal{S} \subset \mathcal{I}$ and a set of new functions $\mathbf{f}^I_{\mathcal{S}} : \mathbf{X}_{\mathcal{S}} \times \mathbf{E}_{\mathcal{J}} \to \mathbf{X}_{\mathcal{S}}$, mechanism surgery "replaces" the corresponding parts of $\mathbf{f}_{\mathcal{I}}$ with $\mathbf{f}^I_{\mathcal{S}}$.

**Definition 7.1.12** (Mechanism surgery)**.** Let $\mathcal{M}$ be the set of SCMs with elements $\langle \mathcal{I}, \mathcal{J}, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{\mathcal{J}}, \_, \mathbb{P}_{\mathcal{E}}, \boldsymbol{E}_{\mathcal{J}} \rangle$ (note that the causal mechanisms are unspecified). Mechanism surgery is an operation $I : \mathbf{X}_{\mathcal{I}}^{\mathbf{X}_{\mathcal{I}} \times \mathbf{E}_{\mathcal{J}}} \times \mathcal{M} \to \mathcal{M}$ that takes a causal model $\mathcal{M}$ with arbitrary causal mechanisms and a set of causal mechanisms $\mathbf{f}'_{\mathcal{I}}$ and maps it to a model $\mathcal{M}' = \langle \mathcal{I}, \mathcal{J}, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{\mathcal{J}}, \mathbf{f}'_{\mathcal{I}}, \mathbb{P}_{\mathcal{E}}, \boldsymbol{E}_{\mathcal{J}} \rangle$.

If $\mathcal{M}$ has causal mechanisms $\mathbf{f}_{\mathcal{I}}$ and $\mathcal{O} \subset \mathcal{I}$ is the largest set such that $\pi_{\mathcal{O}} \circ \mathbf{f}_{\mathcal{I}} = \pi_{\mathcal{O}} \circ \mathbf{f}'_{\mathcal{I}}$ then we say $I$ is an *intervention* on $\mathcal{L} := \mathcal{I} \setminus \mathcal{O}$. We will use the special notation $\mathcal{M}^{I(\mathcal{L}), \mathbf{f}'_{\mathcal{L}}} := I(\mathcal{M}, \mathbf{f}'_{\mathcal{L}}$ to denote an SCM related to $\mathcal{M}$ by intervention on a subset of $\mathcal{I}$.

If *furthermore* $\pi_{\mathcal{L}} \mathbf{f}'_{\mathcal{I}}$ is a constant function equal to $\mathbf{a}$, then we say $I$ is a *hard intervention* on $\mathcal{L}$. We use the special notation $\mathcal{M}^{Do(\mathcal{L})=\mathbf{a}} := I(\mathcal{M}, \mathbf{f}'_{\mathcal{L}}$ to denote SCMs related to $\mathcal{M}$ by hard interventions. We also say that the *causal effect* of $\mathcal{L}$ is the set of SCNMs $\{\mathcal{M}^{Do(\mathcal{L})=\mathbf{a}} | a \in \mathbf{X}_{\mathcal{L}}\}$.

We say a *causal model* is any kind of model that defines causal effects. An SCM $\mathcal{M}$ in combination with hard interventions defines causal effects, so an SCM is a causal model. Call each interventional model $\mathcal{M}^{do(\mathsf{X}_i = x)}$ a *submodel* of $\mathcal{M}$.

Strictly, the random variables $\mathsf{X}_i$ depend on the probability space induced by a particular model $\mathcal{M}$, they are intended to refer to "the same variable" across different models that are related by mechanism surgery. We will abuse notation and use $\mathsf{X}_i$ to refer to the *family* of random variables induced by a set of models related by mechanism surgery, and rely on explicitly noting the measure $\mathbb{P}_{...}(...)$ to specify exactly which random variables we are talking about.

Incidentally, this messiness with random variables can be solved if we use See-Do models.

In practice, we typically specify a "small" SCM containing a few endogenous variables $\mathcal{I}$ (called a "marginal SCM" by Bongers et al. (2016)) which is understood to summarise the relevant characteristics of a "large" SCM containing many variables $\mathcal{I}^*$. We will argue that without restrictions on the large set of variables $\mathcal{I}^*$, surgically transformed SCMs will usually be invalid.

### 7.1.5 Not all variables have well-defined interventions

A long-running controversy about causal inference concerns the question of "the causal effect of body mass index on mortality". On the one hand, Hernán and Taubman (2008) and others claim that there is no well-defined causal effect of a person's body mass index (BMI), defined as their weight divided by their height, and their risk of death. Pearl claims, in defence of Causal Bayesian Networks, that the causal effect of *obesity* is well-defined, though it is not clear whether he defends the proposition that BMI itself has a causal effect:

> That BMI is merely a coarse proxy of obesity is well taken; obesity should ideally be described by a vector of many factors, some are

> easy to measure and others are not. But accessibility to measurement has no bearing on whether the effect of that vector of factors on morbidity is "well defined" or whether the condition of consistency is violated when we fail to specify the interventions used to regulate those factors. (Pearl, 2018)

We argue that BMI does *not* have a well-defined causal effect, and without further assumptions neither does any variable.

### Necessary relationships in cyclic SCMs

If an SCM contains variables that are necessarily related, we wish to impose the additional restriction that these necessary relationships hold for every submodel. This can be done by extending the previous definition:

**Definition 7.1.13** (SCM with necessary relationships)**.** An SCM with necessary relationships (SCNM) is a tuple $\mathcal{M} := \langle \mathcal{I}, \mathcal{J}, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{\mathcal{J}}, \mathbf{f}_{\mathcal{I}}, \mathbf{g}_{\mathcal{I}}, \mathbb{P}_{\mathcal{E}}, \boldsymbol{F}_{\mathcal{J}} \rangle$, which is an SCM with the addition of a vector function of *necessary relationships* $\mathbf{g}_{\mathcal{I}} := \underline{\otimes}_{i \in \mathcal{I}} g_i$ where each $g_i : \mathbf{X}_{\mathcal{I}} \to X_i$ is a necessary relationship involving $\mathsf{X}_i$.

An SCM with necessary induces a unique probability space if for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$ there exists a unique $\mathbf{X} \in \mathbf{X}_{\mathcal{I}}$ such that simultaneously

$$\mathbf{X} = \mathbf{f}_{\mathcal{I}}(\mathbf{X}, \mathbf{e}) \tag{7.45}$$

$$\mathbf{X} = \mathbf{g}_{\mathcal{I}}(\mathbf{X}) \tag{7.46}$$

If no such $\mathbf{X}$ exists then an SCNM is invalid.

Mechanism surgery for SCNMs involves modification of $\mathbf{f}_{\mathcal{I}}$ only, just like SCMs.

If we wish to stipulate that a particular variable $\mathsf{X}_i$ has no causal relationships or necessary relationships we can specify this with the trivial mechanisms $f_i : (\mathbf{X}, \mathbf{e}) \mapsto x_i$ and $g_i : \mathbf{X} \mapsto x_i$ respectively. An SCNM $\mathcal{M}$ with the trivial necessary relationship $\mathbf{g}_{\mathcal{I}} : \mathbf{X} \mapsto \mathbf{X}$ induces the equivalent probability spaces as the SCM obtained by removing $\mathbf{g}_{\mathcal{I}}$ from $\mathcal{M}$.

Because BMI is always equal height/weight, given some SCNM $\mathcal{M}$ containing endogenous variables $\mathsf{X}_h$, $\mathsf{X}_w$ and $\mathsf{X}_b$ representing height, weight and BMI respectively, it should be possible to construct a more "primitive" SCNM $\mathcal{M}^p$ containing every variable $\mathcal{M}$ does except $\mathsf{X}_b$ that agrees with $\mathcal{M}$ on all interventions except those on $\mathsf{X}_b$.

**Definition 7.1.14** (Marginal model)**.** Given an SCNM

$$\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{\mathcal{J}}, \mathbf{f}_{\mathcal{I}}, \mathbf{g}_{\mathcal{I}}, \mathbb{P}_{\mathcal{E}}, \boldsymbol{F}_{\mathcal{J}} \rangle$$

a marginal model over $\mathcal{L} \subset \mathcal{I}$ is an SCNM

$$\mathcal{M}^{\mathbin{\text{⫽}}_{\mathcal{L}}} = \langle \mathcal{O}, \mathcal{J}, \mathbf{X}_{\mathcal{O}}, \mathbf{E}_{\mathcal{J}}, \mathbf{f}_{\mathcal{O}}^{\mathcal{L}*}, \mathbf{g}_{\mathcal{O}}^{\mathcal{L}*}, \mathbb{P}_{\mathcal{E}}, \boldsymbol{F}_{\mathcal{J}} \rangle$$

such that $(\mathbb{P}_{\mathcal{M}})^{\blacklozenge}_{\mathcal{L}} = \mathbb{P}_{(\mathcal{M}^{\blacklozenge}{}_{L})}$ and for all interventions $\mathbf{f}'_{\mathcal{O}}$ on $\mathcal{O} := \mathcal{I} \setminus \mathcal{L}$ that do not depend on $\mathcal{L}$

$$(\mathbb{P}_{\mathcal{M}^{I(\mathcal{O}),\mathbf{f}'_{\mathcal{O}}}})^{\blacklozenge}_{\mathcal{L}} = \mathbb{P}_{(\mathcal{M}^{\blacklozenge}{}_{\mathcal{L}},I(\mathcal{O}),\mathbf{f}'_{\mathcal{O}}\circ\pi_{\mathcal{O}})}$$

A *primitive model* is a special case of a marginal model where any intervention that depended only on endogenous variables in the original model can be replicated with some intervention that depends only on endogenous variables in the marginal model. If the endogenous variables represent *observed* variables, then the plausible intervention operations may only be allowed to depend on these variables. In general, there may be interventions that are possible in the original model that are not possible in the marginal model.

**Definition 7.1.15** (Primitive model)**.** A *primitive model* $\mathcal{M}^p$ is a marginal model of $\mathcal{M}$ with respect to some $\mathcal{L}$ such that for all interventions $\mathbf{f}'_{\mathcal{O}}$ that do not depend on $\mathcal{J}$ there exists some $\mathbf{g}'_{\mathcal{O}} : \mathbf{X}_{\mathcal{O}} \times \mathbf{E}_{\mathcal{J}} \to \mathbf{X}_{\mathcal{O}}$ that does not depend on $\mathcal{J}$ such that

$$(\mathbb{P}_{\mathcal{M}^{I(\mathcal{O}),\mathbf{f}'_{\mathcal{O}}}})^{\blacklozenge}_{\mathcal{L}} = \mathbb{P}_{(\mathcal{M}^{\blacklozenge}{}_{\mathcal{L}},I(\mathcal{O}),\mathbf{g}'_{\mathcal{O}})}$$

We claim that given any SCNM $\mathcal{M}$ containing endogenous variables $\mathsf{X}_h$, $\mathsf{X}_w$ and $\mathsf{X}_b$ representing height, weight and BMI there should be a primitive model $\mathcal{M}^p$ of $\mathcal{M}$ with respect to $\{p\}$.

**Lemma 7.1.16** (Primitive models)**.** *$\mathcal{M}^p$ is a primitive model of $\mathcal{M}$ with respect to $\mathcal{L} \subset \mathcal{I}$ iff $S(\pi_{\mathcal{O}}\mathbf{f}_{\mathcal{I}}) \overset{a.s.}{=} S(\mathbf{f}^p_{\mathcal{O}})$ for $\mathcal{O} := \mathcal{I} \setminus \mathcal{L}$ and for all $\mathbf{X} \in \mathbf{X}_{\mathcal{I}}$, $\mathbf{g}$*

However, as Theroem 7.1.18 shows, if an SCNM with height, weight and BMI can be derived from an SCNM containing just height and weight then there are no valid hard interventions on BMI.

**Definition 7.1.17** (Derived model)**.** Given a SCNM $\mathcal{M} := \langle \mathcal{I}, \mathcal{J}, \mathbf{X}_{\mathcal{I}}, \mathbf{E}_{\mathcal{J}}, \mathbf{f}_{\mathcal{I}}, \mathbf{g}_{\mathcal{I}}, \mathbb{P}_{\mathcal{E}}, \boldsymbol{F}_{\mathcal{J}} \rangle$, say $\mathcal{M}' = \langle \mathcal{I}', \mathcal{J}, \mathbf{X}_{\mathcal{I}'}, \mathbf{E}_{\mathcal{J}}, \mathbf{f}'_{\mathcal{I}'}, \mathbf{g}'_{\mathcal{I}'}, \mathbb{P}_{\mathcal{E}}, \boldsymbol{F}_{\mathcal{J}} \rangle$ is *derived* from $\mathcal{M}$ if there exists some additional index/variable/relationships $i' \notin \mathcal{I}, X_{i'}$ such that

$$\mathcal{I}' = \mathcal{I} \cup \{i'\} \tag{7.47}$$

$$\mathbf{X}_{\mathcal{I}'} = \mathbf{X}_{\mathcal{I}} \cup X_{i'} \tag{7.48}$$

and, defining $\pi_{\mathcal{I}'\setminus i'} : \mathbf{X}_{\mathcal{I}'} \to \mathbf{X}_{\mathcal{I}}$ as the projection map that "forgets" $\mathsf{X}_{i'}$, for any $\mathbf{e} \in \mathbf{E}_{\mathcal{J}}$ we have

$$\mathbf{X}' = \mathbf{f}'_{\mathcal{I}'}(\mathbf{X}', \mathbf{e}) \tag{7.49}$$

$$\text{and } \mathbf{X}' = \mathbf{g}'_{\mathcal{I}'}(\mathbf{X}') \implies \pi_{\mathcal{I}'\setminus i'}(\mathbf{X}') = \mathbf{f}_{\mathcal{I}}(\pi_{\mathcal{I}'\setminus i'}(\mathbf{X}'), \mathbf{e}) \tag{7.50}$$

$$\text{and } \pi_{\mathcal{I}'\setminus i'}(\mathbf{X}') = \mathbf{g}'_{\mathcal{I}'}(\pi_{\mathcal{I}'\setminus i'}(\mathbf{X}')) \tag{7.51}$$

**Theorem 7.1.18** (Interventions and necessary relationships don't mix)**.** *If $\mathcal{M}'$ is derived from $\mathcal{M}$ with the additional elements $i', X_{i'}, f_{i'}, g_{i'}$ and both $\mathcal{M}$ and $\mathcal{M}'$ are uniquely solvable and $\mathbb{P}_{\mathcal{X}'\otimes\mathcal{E}}(\mathsf{X}_{i'})$ is not single valued then no hard interventions on $\mathsf{X}_{i'}$ are possible.*

*Proof.* Because $\mathcal{M}$ is uniquely solvable, for $\mathbb{P}_{\mathcal{E}}$ almost every $\mathbf{e}$ there is a unique $\mathbf{X}^e$ such that

$$\mathbf{X}^e = \mathbf{f}_{\mathcal{I}}(\mathbf{X}^e, \mathbf{e}) \tag{7.52}$$
$$\mathbf{X}^e = \mathbf{g}_{\mathcal{I}}(\mathbf{X}^e) \tag{7.53}$$

Because $\mathcal{M}'$ is also uniquely solvable, for $\mathbb{P}_{\mathcal{E}}$ almost every $\mathbf{e}$ we have $\mathbf{X}'^e \in \mathbf{X}_{\mathcal{I}'}$ such that $\pi_{\mathcal{I}'\setminus i'}(\mathbf{X}')^{\prime e} = \mathbf{X}^e$ and

$$x'^e_{i'} = \mathbf{g}_{i'}(\mathbf{X}'^e) \tag{7.54}$$

Because $\mathbb{P}_{\mathcal{X}'\otimes\mathcal{E}}(\mathsf{X}_{i'})$ is not single valued there are non-null sets $A, B \in \mathcal{E}$ such that $e_a \in A$, $e_b \in B$ implies

$$\mathbf{g}_{i'}(\mathbf{X}'^{e_a}) \neq \mathbf{g}_{i'}(\mathbf{X}'^{e_b}) \tag{7.55}$$

Therefore there exists no $a \in X_{i'}$ that can simultaneously satisfy 7.54 for almost every $\mathbf{e}$. However, any hard intervention $\mathcal{M}'^{,do(\mathsf{X}_{i'}=a)}$ requires such an $a$ in order to be solvable. □

**Corollary 7.1.19.** *Either there are no hard interventions defined on BMI or there is no SCNM containing height and weight with a unique solution from which an SCNM containing height, weight and BMI can be derived.*

I can formalise the following, but I'm just writing it out so I can get to the end for now

The problem posed by Theorem 7.1.18 can be circumvented to some extent by joint interventions. Consider the variables $\mathsf{X}_1$ and $\mathsf{X}_2$ where $\mathsf{X}_1 = -\mathsf{X}_2$ necessarily. While Theorem 7.1.18 disallows interventions on $\mathsf{X}_2$ individually (supposing we can obtain a unique model featuring only $\mathsf{X}_1$), it does not disallow interventions that jointly set $\mathsf{X}_1$ and $\mathsf{X}_2$ to permissible values. In this case, this is unproblematic as the only joint intervention that sets $\mathsf{X}_1$ to 1 must also set $\mathsf{X}_2$ to $-1$.

If we have non-invertible necessary relationships such as $\mathsf{X}_1 = \mathsf{X}_2 + \mathsf{X}_3$, however, there are now *multiple* joint interventions on $\mathsf{X}_1$ that can be performed. I regard this as the most plausible solution to the difficulties raised so far: for variables that are in non-invertible necessary relationships, there is a set of operations associated with the "intervention" that sets $\mathsf{X}_1 = 1$.

However, we still need to make sure the interventions that we have supposed comprise the operations associated with setting $\mathsf{X}_1 = 1$ exist themselves. It is sufficient that the SCNM with $\mathsf{X}_1$ is derived from a higher order *uniquely solvable SCM* with $\mathsf{X}_2$ and $\mathsf{X}_3$ only .

And necessary? There might be "degenerate" necessary relationships that don't harm the possibility of defining interventions, and I'd need to show an equivalence to an SCM in this case

because interventions are defined in uniquely solvable SCMs and derivation preserves interventions on the old variables

If any variables are included in a causal model that are necessarily related to other variables (and honestly, is there any variable that isn't?), it is not enough to suppose that the model being used is a marginalisation of some larger causal model. Rather, it must be obtained by derivation and marginalisation from some model that represents the basic interventions that are possible, which we call the *atomic model*.

**References**

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.

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.

Florence Clerc, Fredrik Dahlqvist, Vincent Danos, and Ilias Garnier. Pointless learning. *20th International Conference on Foundations of Software Science and Computation Structures (FoSSaCS 2017)*, March 2017. doi: 10.1007/978-3-662-54458-7_21. URL https://www.research.ed.ac.uk/portal/en/publications/pointless-learning(694fb610-69c5-469c-9793-825df4f8ddec).html.

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, 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.

Thomas S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, July 1967. ISBN 978-1-4832-2123-6.

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, 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.

D. Heckerman and R. Shachter. Decision-Theoretic Foundations for Causal Reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, December 1995. ISSN 1076-9757. doi: 10.1613/jair.202. URL https://www.jair.org/index.php/jair/article/view/10151.

M. A. Hernán and S. L. Taubman. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity*, 32(S3):S8–S14, August 2008. ISSN 1476-5497. doi: 10.1038/ijo.2008.82. URL https://www.nature.com/articles/ijo200882.

Edwin Hewitt and Leonard J. Savage. Symmetric Measures on Cartesian Products. *Transactions of the American Mathematical Society*, 80(2):470–501, 1955. ISSN 0002-9947. doi: 10.2307/1992999. URL https://www.jstor.org/stable/1992999. Publisher: American Mathematical Society.

Alan Hájek. Interpretations of Probability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019. URL https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/.

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.

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 Publications Sage 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.

Graham Loomes and Robert Sugden. Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *The Economic Journal*, 92(368):805–824, 1982. ISSN 0013-0133. doi: 10.2307/2232669. URL https://www.jstor.org/stable/2232669.

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. Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes. *Journal of Causal Inference*, 6(2), 2018. doi: 10.1515/jci-2018-2001. URL https://www.degruyter.com/view/j/jci.2018.6.issue-2/jci-2018-2001/jci-2018-2001.xml.

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.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5): 947–1012, 2016. ISSN 1467-9868. doi: 10.1111/rssb.12167. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12167.

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.

Donald B. Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, March 2005. ISSN 0162-1459. doi: 10.1198/016214504000001880. URL https://doi.org/10.1198/016214504000001880.

Leonard J. Savage. *The Foundations of Statistics*. Wiley Publications in Statistics, 1954.

Peter Selinger. A survey of graphical languages for monoidal categories. *arXiv:0908.3347 [math]*, 813:289–355, 2010. doi: 10.1007/978-3-642-12821-9_4. URL http://arxiv.org/abs/0908.3347. arXiv: 0908.3347.

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.

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.

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: