Causal Statistical Decision Theory|What are interventions?

David Johnston

June 25, 2021

Contents

0.1 Theories of causal inference

Feedback start here

Beginning in the 1930s, a number of associations between cigarette smoking and lung cancer were established: on a population level, lung cancer rates rose rapidly alongside the prevalence of cigarette smoking. Lung cancer patients were far more likely to have a smoking history than demographically similar individuals without cancer and smokers were around 40 times as likely as demographically similar non-smokers to go on to develop lung cancer. In laborotory experiments, cells which were introduced to tobacco smoke developed *ciliastasis*, and mice exposed to cigarette smoke tars developed tumors(Proctor, 2012). Nevertheless, until the late 1950s, substantial controversy persisted over the question of whether the available data was sufficient to establish that smoking cigarettes *caused* lung cancer. Cigarette manufacturers famously argued against

any possible connection (Oreskes and Conway, 2011) and Roland Fisher in particular argued that the available data was not enough to establish that smoking actually caused lung cancer (Fisher, 1958). Today, it is widely accepted that cigarettes do cause lung cancer, along with other serious conditions such as vascular disease and chronic respiratory disease (World Health Organisation, 2018; Wiblin, 2016).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A key feature of CSDT is what I call the *option set*. This is the set of decisions, acts or counterfactual propositions under consideration in a given

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

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

[

Todo: I need the following theorem in this chapter]

Theorem 0.1.1 (Representation)**.**

Representation theorem: can uniquely define kernel $P^{\mathsf{X}|\mathsf{Y}}$ *with* $P^{\mathsf{Z}|\mathsf{Y}}$ *and* $P^{\mathsf{X}|\mathsf{ZY}}$

Todo: conditional expectation, martingale convergence

0.1.1 Probability Theory

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

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

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

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

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

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

Let T be all the open subsets of $\mathbb{R}$. Then $\mathcal{B}(\mathbb{R}) := \sigma(T)$ is the *Borel σ-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 $x \in X$, $A \in \mathcal{Y}$

$$\mathbb{M}_x(A) = \mathbb{N}_x(A) \tag{1}$$

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

0.1.2 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 $A \in \mathcal{Y}$,

$$\mu\mathbb{M}(A) := \int_X \mathbb{M}_x(A) d\mu(x) \tag{2}$$

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

$$\mathbb{M}\mathsf{T}(x) := \int_Y T(y) d\mathbb{M}_x(y) \tag{3}$$

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

$$(\mathbb{M}\mathbb{N})_x(A) := \int_Y \mathbb{N}_y(A) d\mathbb{M}_x(y) \tag{4}$$

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 $H \in \mathcal{Z}$:

$$(\mathbb{M} \otimes \mathbb{N})_{x,y}(G \times H) := \mathbb{M}_x(G)\mathbb{N}_y(H) \tag{5}$$

0.1.3 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:

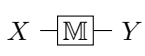
- String diagrams as they are used in this work can always be interpreted as a mixture of kernel-kernel products and tensor products of Markov kernels
- String diagrams are the subject of a coherence theorem: two string diagrams that differ only by planar deformation are always equal (Selinger, 2010). 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 μ' can be given).

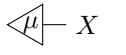
A Markov kernel $\mathbb{M} : X \to \Delta(\mathcal{Y})$ can always be represented as a rectangular box with input and output wires labeled with the relevant spaces:



(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:



(7)

We can exploit the identification of the probability measure μ 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.

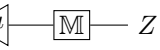
Cho and Jacobs (2019) defines the operation of *conditioning* using kernel-function products which makes use of kernel-function products which and, unlike measure-kernel products, kernel-function products do not in general produce Markov kernels. At this stage, we do not make use of a graphical conditioning 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

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

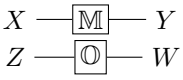
(8)

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

$\mu\mathbb{M} :=$ 

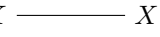
(9)

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

$\mathbb{M} \otimes \mathbb{O} :=$ 

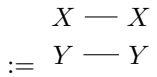
(10)

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:

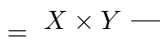
$\mathrm{Id}^X :=$ 

(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:

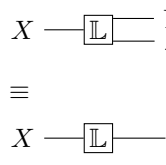
$X \times Y \cong \mathrm{Id}^X \otimes \mathrm{Id}^Y :=$ 

(12)

$=$ 

(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:



(14)

$\equiv$ (15)

(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 δ_x that places all mass on the same point:

$$\mathbf{Id} : x \mapsto \delta_x \equiv \quad X \text{ — } X \tag{17}$$

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

$$\begin{aligned}
(\mathbb{K}\mathbf{Id})_w(A) &= \int_X \mathbf{Id}_x(A) d\mathbb{K}_w(x) && (18)\\
&= \int_X \delta_x(A) d\mathbb{K}_w(x) && (19)\\
&= \int_A d\mathbb{K}_w(x) && (20)\\
&= \mathbb{K}_w(A) && (21)\\
(\mathbf{Id}\mathbb{K})_w(A) &= \int_X \mathbb{K}_x(A) d\mathbf{Id}_w(x) && (22)\\
&= \int_X \mathbb{K}_x(A) d\delta_w(x) && (23)\\
&= \mathbb{K}_w(A) && (24)
\end{aligned}$$

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

$$\curlyvee : x \mapsto \delta_{(x,x)} \equiv \quad X \text{ —}\!\!\left\langle \begin{matrix} X \\ X \end{matrix}\right. \tag{25}$$

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

$$\begin{aligned}
\int_(X \times X) \mathbb{K}_{x',x''}(A) d\curlyvee_x(x', x'') &= \int_(X \times X) \mathbb{K}_{x',x''}(A) d\delta_{(x,x)}(x', x'') && (26)\\
&= \mathbb{K}_{x,x}(A) && (27)
\end{aligned}$$

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

$$\rho := (x, y) \to \delta_{(y,x)} \equiv \begin{matrix} Y \\ X \end{matrix} \asymp \begin{matrix} X \\ Y \end{matrix} \tag{28}$$

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

$$(\rho\mathbb{K})_{y,x}(A) = \int_{(} X \times Y)\mathbb{K}_{x',y'}(A)d\rho_{(y,x)}(x', y') \quad = \int_{(} X \times Y)\mathbb{K}_{x',y'}(A)d\delta_{(x,y)}(x', y') \tag{29}$$

$$= \mathbb{K}_{x,y}(A) \tag{30}$$

$$(\mathbb{L}\rho)_z(B \times A) = \int_{X \times Y} \rho_{x',y'}(B \times A)d\mathbb{L}_z(x', y') \tag{31}$$

$$= \int_{X \times Y} \delta_{(y',x')}(B \times A)d\mathbb{L}_z(x', y') \tag{32}$$

$$= \mathbb{L}_z(A \times B) \tag{33}$$

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

$$* : x \mapsto \delta_* \equiv \begin{matrix} X \\ \end{matrix} \longrightarrow * \tag{34}$$

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 $\mu \in \Delta(\mathcal{W})$, μg 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 0.1.2 (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 0.1.3 (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 0.1.4 (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 $a \in V$, $B \in \mathcal{X}$:

$$(\mathbb{M}\mathbb{F}^g)_a(B) = \int_W \delta_{g(y)}(B) d\mathbb{M}_a(y) \tag{35}$$

$$= \int_W \delta_y(g^{-1}(B)) d\mathbb{M}_a(y) \tag{36}$$

$$= \int_{g^{-1}(B)} d\mathbb{M}_a(y) \tag{37}$$

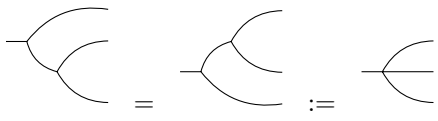
$$= (g_\# \mathbb{M})_a(B) \tag{38}$$

□

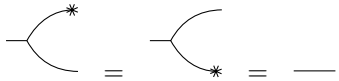
Working With String Diagrams

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.

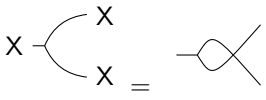
Axioms of Symmetric Monoidal Categories For the following, we either omit labels or label diagrams with their domain and codomain spaces, as we are discussing identities of kernels rather than identities of components of a condtional probability space. Recalling the unique Markov kernels defined above,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



(39)

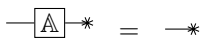


(40)



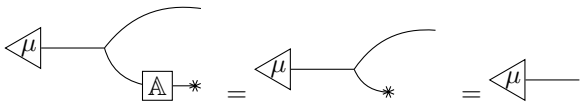
(41)

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



(42)

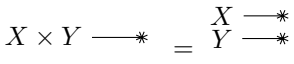
Combining 40 and 42 we can derive the following: integrating $\mathbb{A} : X \to \Delta(\mathcal{Y})$ with respect to $\mu \in \Delta(\mathcal{X})$ and then discarding the output of $\mathbb{A}$ leaves us with μ:



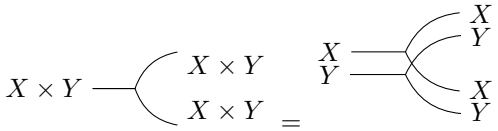
(43)

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 $\curlyvee$:

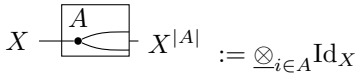


(44)



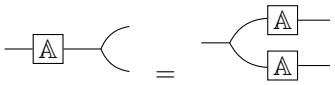
(45)

Note that for some set X, the copy map $\curlyvee_X = \mathrm{Id}_X \underline{\otimes} \mathrm{Id}_X$. This combined with 41 allows us to define the A-copy map for some $A \subseteq \mathbb{N}$:



(46)

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



(47)

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

Definition 0.1.5 (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 0.1.6 (Copy map commutes for deterministic kernels (Fong, 2013))**.** *Equation 47 holds iff $\mathbb{A}$ is deterministic.*

Examples

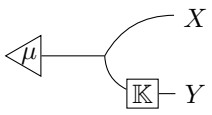
Given $\mu \in \Delta(X), \mathbb{K} : X \to \Delta(Y)$, $A \in \mathcal{X}$ and $B \in \mathcal{Y}$:

$$A \times B \mapsto \int_A \mathbb{K}(x; B) d\mu(x) \tag{48}$$

$$\equiv \tag{49}$$

$$\mu \curlyvee (\mathbf{Id}_X \otimes \mathbb{K}) \tag{50}$$

$$\equiv \tag{51}$$



(52)

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

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 ν^{Y} is obtained by marginalising over "X" (a notion that can be made more precise by assigning names to wires). Then

$$\nu(* \otimes \mathrm{Id}^Y) = \tag{53}$$

$$\nu(* \otimes \mathrm{Id}^Y)(B) := \nu(* \otimes \mathrm{Id}^Y)(B \times \{*\}) \tag{54}$$

$$= \int_{X \times Y} \mathrm{Id}^Y_y(B) *_x (\{*\}) d\nu(x, y) \tag{55}$$

$$= \int_{X \times Y} \delta_y(B) \delta_*(\{*\}) d\nu(x, y) \tag{56}$$

$$= \int_{X \times B} d\nu(x, y) \tag{57}$$

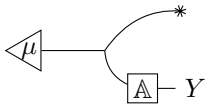
$$= \nu(X \times B) \tag{58}$$

$$= \nu^{\mathsf{Y}}(B) \tag{59}$$

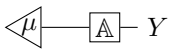
Thus the action of the erasing wire "X" is equivalent to marginalising over "X".

Consider the result of marginalising 52 over "X":

$$\nu^Y(B) = \tag{60}$$



$$= \tag{61}$$



0.1.4 Random Variables

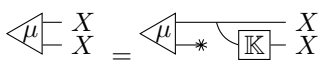
The summary of this section is:

- 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.

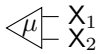
We use the same approach here, with a couple of additions. We are interested in variables whose outcomes depend both on random processes and decisions. Suppose that given a particular distribution over decision variables, a probability distribution over the decision variables and random variables is 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:



(62)

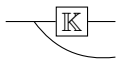
Suppose that we also assign the names X_1 to the upper output wire and X_2 to the lower output wire in the diagram of μ:



(63)

Then it seems sensible to call $\mathbb{K}$ "the probability of X_2 given 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 0.1.7 (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

$\mathbb{K}^* :=$ 

(64)

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

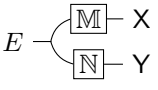
A *probability space* $(\mathbb{P}, \Omega, \mathcal{F})$ is a probability measure $\mathbb{P} : \Delta(\Omega)$, which we call the *ambient measure*, along with the *sample space* Ω and the *events* $\mathcal{F}$. A probability space is equivalent to a Markov kernel space with domain $D = \{*\}$ - note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

Definition 0.1.8 (Random variable)**.** Given a Markov kernel space $(\mathbb{K}, D, \Omega)$, a random variable X is a measurable function $\Omega \times D \to E$ for arbitrary measurable E.

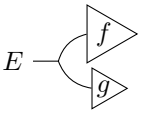
Definition 0.1.9 (Domain variable)**.** Given a Markov kernel space $(\mathbb{K}, D, \Omega)$, the *domain variable* $\mathsf{D} : \Omega \times D \to D$ is the distinguished random variable $\mathsf{D} : (x, d) \mapsto d$.

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 0.1.10 (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:

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

(65)

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

(66)

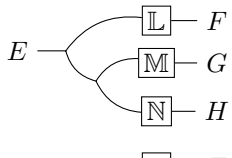
The operation denoted by $\underline{\otimes}$ is associative (Lemma 0.1.11), so we can without ambiguity write $f \underline{\otimes} g \underline{\otimes} h = (f \underline{\otimes} g) \underline{\otimes} h = f \underline{\otimes} (g \underline{\otimes} h)$ for finite groups of functions or Markov 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 0.1.11

Lemma 0.1.11 ($\underline{\otimes}$ is associative)**.** *For Markov kernels* $\mathbb{L} : E \to \delta(\mathcal{F})$, $\mathbb{M} : E \to \delta(\mathcal{G})$ *and* $\mathbb{N} : E \to \delta(\mathcal{H})$, $(\mathbb{L} \underline{\otimes} \mathbb{M}) \underline{\otimes} \mathbb{N} = \mathbb{L} \underline{\otimes} (\mathbb{M} \underline{\otimes} \mathbb{N})$.

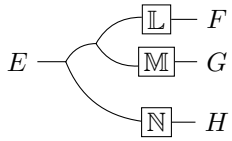
Proof.

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



(67)

$=$



(68)

$$= (\mathbb{L} \underline{\otimes} \mathbb{M}) \underline{\otimes} \mathbb{N} \tag{69}$$

This follows directly from Equation 39. □

Definition 0.1.12 (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 X is the probability measure $\mathbb{P}^{\mathsf{X}} := \mathbb{P}\mathbb{F}^{\mathsf{X}}$.

See Lemma 0.1.4 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 0.1.7)

Definition 0.1.13 (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 X and 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 0.1.14 which follows.

Lemma 0.1.14 (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.*

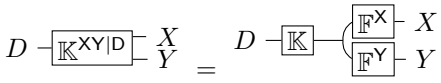
Given $\mathsf{X} : \Omega \to G$ *and* $\mathsf{Y} : \Omega \to H$, $\mathbb{F}^{\mathsf{X} \underline{\otimes} \mathsf{Y}} = \mathbb{F}^{\mathsf{X}} \underline{\otimes} \mathbb{F}^{\mathsf{Y}}$

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

$$\begin{aligned}
\mathbb{F}^{\mathsf{X} \underline{\otimes} \mathsf{Y}}(a; B \times C) &= \delta_{\mathsf{X}(a), \mathsf{Y}(a)}(B \times C) && (70)\\
&= \delta_{\mathsf{X}(a)}(B) \delta_{\mathsf{Y}(a)}(C) && (71)\\
&= (\delta_{\mathsf{X}(a)} \otimes \delta_{\mathsf{Y}(a)})(B \times C) && (72)\\
&= \mathbb{F}^{\mathsf{X}} \underline{\otimes} \mathbb{F}^{\mathsf{Y}} && (73)
\end{aligned}$$

Equality follows from the monotone class theorem. □

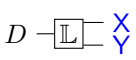
Corollary 0.1.15. *Given a Markov kernel space* $(\mathbb{K}, \Omega, D)$ *and random variables* $\mathsf{X} : \Omega \times D \to X$, $\mathsf{Y} : \Omega \times D \to Y$, *the following holds:*



(74)

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

Definition 0.1.16 (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 $\mathbf{L}$:

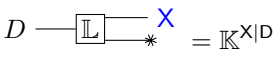


(75)

is *valid* iff

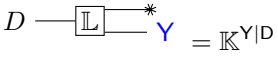
$$\mathbb{L} = \mathbb{K}_{\mathsf{XY|D}} \tag{76}$$

and



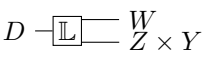
(77)

and



(78)

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

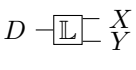


(79)

For *this* diagram, properties 77 and 78 do not hold, even though 76 does.

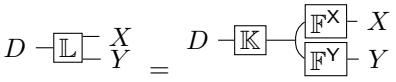
Lemma 0.1.17 (Output label assignments exist)**.** *Given Markov kernel space* $(\mathbb{K}, D, \Omega)$*, random variables* $\mathsf{X} : \Omega \times D \to X$ *and* $\mathsf{Y} : \Omega \times D \to Y$ *then there exists a diagram of* $\mathbb{L} := \mathbb{K}^{\mathsf{XY|D}}$ *with a valid output labelling assigning* X *and* 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



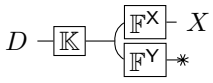
(80)

By Corollary 0.1.15, we have



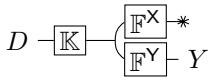
(81)

Therefore



$$= \mathbb{K}\mathbb{F}^{\mathsf{X}} \tag{82}$$

$$= \mathbb{K}^{\mathsf{X}|\mathsf{D}} \tag{83}$$



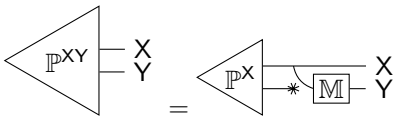
$$= \mathbb{K}\mathbb{F}^{\mathsf{Y}} \tag{84}$$

$$= \mathbb{K}^{\mathsf{Y}|\mathsf{D}} \tag{85}$$

□

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

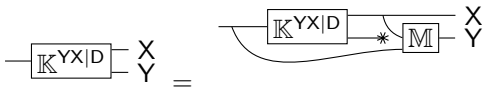
Definition 0.1.18 (Disintegration)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$, and random variables X and Y, we say that $\mathbb{M} : E \to \Delta(\mathcal{F})$ is a Y *given* X *disintegration* of $\mathbb{P}$ iff



(86)

$\mathbb{M}$ is a version of $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$, "the probability of Y given X". Let $\mathbb{P}^{\{\mathsf{Y}|\mathsf{X}\}}$ be the set of all kernels that satisfy 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 Y *given* DX *disintegration* of $\mathbb{K}^{\mathsf{YX}|\mathsf{D}}$ iff



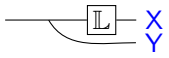
(87)

Write $\mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$ for the set of kernels satisfying 87 and $\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}$ for an arbitrary member of $\mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$.

Definition 0.1.19 (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.

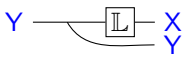
If an input wire is connected to an output wire and that output wire has a valid label X, then it is valid to label the input wire with X.

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



(88)

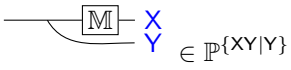
i.e. if $\mathbb{L} \in \mathbb{P}^{\{\mathsf{XY}|\mathsf{Y}\}}$, then the following is a valid input label:



(89)

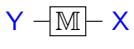
An input wire in a diagram for $\mathbb{M}$ may be labeled X *if and only if* copy and identity maps can be inserted to yield a diagram in which the input wire labeled X is connected to an output wire with valid label X.

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



(90)

and hence the output labels are valid. Diagram 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

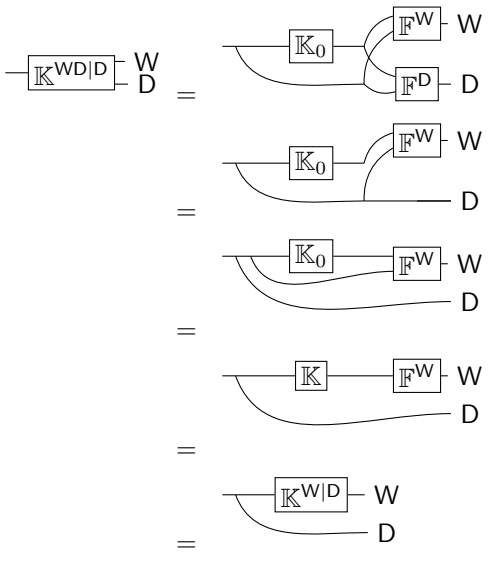


(91)

Lemma 0.1.20 (Labeling of disintegrations)**.** *Given a kernel space* $(\mathbb{K}, D, \Omega)$*, random variables* X *and* Y*, domain variable* D *and disintegration* $\mathbb{L} \in \mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$*, there is a diagram of* $\mathbb{L}$ *with valid input labels* X *and* D *and valid output label* 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 $\mathbb{K}$:

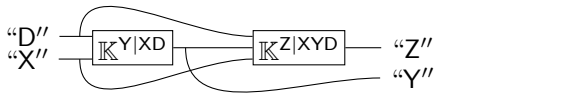


(92) (93) (94) (95) (96)

□

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

Theorem 0.1.21 (Iterated disintegration)**.** *Given a kernel space* $(\mathbb{K}, D, \Omega)$*, random variables* X*,* Y *and* Z *and domain variable* D*,*



(97)

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

$$(d, x; A, B) \mapsto \int_A \mathbb{K}^{\mathsf{Z}|\mathsf{XYD}}_{(x,y,d)}(B) d\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}_{(x,d)}(y) \in \mathbb{K}^{\{\mathsf{ZY}|\mathsf{XD}\}} \tag{98}$$

Proof.

standard result, to write

□

Existence of Disintegrations

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

Theorem 0.1.22 (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 0.1.22, 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 λ-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

$$Q(d)_x(A) = \left(1 - \mathbb{1}_C(x)\mathbb{1}_{\{x\}}(d)\right) \lambda(A) + \mathbb{1}_C(x)\mathbb{1}_{\{x\}}(d)\delta_0(A) \tag{99}$$

That is, Q is the measure λA 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 λ. Thus $Q(d)$ is a version of $(\mathbb{K}_d)^{\{\mathsf{Y}|\mathsf{X}\}}$. Consider the function

$$Q^{\{0\}} : (d, x) \mapsto Q(d)_x(\{0\}) \tag{100}$$

$$\left(Q^{\{0\}}\right)^{-1}(\{1\}) = \{(d, x) : Q(d)_x(\{0\}) = 1\} \tag{101}$$

$$= \{(d, x) : d = x \ \ \& \ \ d \in C\} \tag{102}$$

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 99. 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 0.1.21 to find further disintegrations (see Lemma 0.1.30. 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 δ_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 X and 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 99 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

$$\mathbb{P}^{\mathsf{XY}|\mathsf{X}}_{(\mathsf{X}=x),x'}(A \times B) = \mathbb{1}_C(x)\delta_x(A)\delta_1(B) + (1 - \mathbb{1}_C(x))\delta_x(A)\delta_0(B) \tag{103}$$

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 $x \in X$, 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

$$\mathbb{P}^{\mathsf{XY}|\mathsf{X}}_{(\mathsf{X}=x),x'}(A \times B) = \mathbb{1}_R(x)\delta_{x'}(A)\delta_1(B) + (1 - \mathbb{1}_R(x))\delta_x(A)\lambda(B) \tag{104}$$

While Equation 104 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 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 0.1.23 (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 0.1.24 (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*

$$d \mapsto \mathbb{K}_d(A) \tag{105}$$

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

$$R_d^{A,n}(x) = \sum_{H_i \in \mathcal{J}_n} \mathbb{1}_{H_i}(x) \frac{\mathbb{K}_d^{\mathsf{XY}}(H_i \times A)}{\mathbb{K}_d^{\mathsf{XY}}(H_i \times Y)} \tag{106}$$

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

$$\mathbb{E}[R_d^{A,n}(x)] = \sum_i \mathbb{K}_d^{\mathsf{XY}}(H_i \times A) \tag{107}$$

$$= \mathbb{K}_d^{\mathsf{XY}}(X \times A) \tag{108}$$

$$< \infty \tag{109}$$

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

$$\mathbb{E}[\mathbb{1}_{H_j} R_d^{A,n}] = \int \mathbb{1}_{H_j}(x) \sum_i \mathbb{1}_{H_i}(x) \frac{\mathbb{K}_d^{\mathsf{XY}}(H_i \times A)}{\mathbb{K}_d^{\mathsf{XY}}(H_i \times Y)} d\mathbb{K}_d^{\mathsf{XY}}(x,y) \tag{110}$$

$$= \mathbb{K}_d^{\mathsf{XY}}(H_j \times A) \tag{111}$$

$$= \int \mathbb{1}_{H_j}(x) \sum_i \mathbb{1}_{H_i^{n+1}}(x) \frac{\mathbb{K}_d^{\mathsf{XY}}(H_i^{n+1} \times A)}{\mathbb{K}_d^{\mathsf{XY}}(H_i^{n+1} \times Y)} d\mathbb{K}_d^{\mathsf{XY}}(x,y) \tag{112}$$

$$= \mathbb{E}[\mathbb{1}_{H_j} R_d^{A,n+1}] \tag{113}$$

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 0.1.27), so it converges to a measurable function R_d^A almost surely and also in L^1.

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

$$\int_B R_d^A(x) d\mathbb{K}_d^{\mathsf{XY}}(\{x\} \otimes Y) = \lim_{n\to\infty} \int_B R_d^{A,n}(x) d\mathbb{K}_d^{\mathsf{XY}}(\{x\} \otimes Y) \tag{114}$$

$$\tag{115}$$

By Equation 111, $\int_{H_j} R_d^A(x) d\mathbb{K}_d^{\mathsf{XY}}(\{x\} \otimes Y) = \mathbb{K}_d^{\mathsf{XY}}(H_j \times A)$ 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 δ such that $|d - d'| < \delta$ implies:

$$|R_d^{A,n}(x) - R_{d'}^{A,n}(x)| \leq \left| \frac{(1+\epsilon)\mathbb{K}_d^{\mathsf{XY}}(H(x) \times A) - \mathbb{K}_d^{\mathsf{XY}}(H(x) \times A)(1-\epsilon)}{\mathbb{K}_d^{\mathsf{XY}}(H(x) \times Y)(1+\epsilon)} \right| \tag{116}$$

$$= \left| \frac{2\epsilon \mathbb{K}_d^{\mathsf{XY}}(H(x) \times A)}{\mathbb{K}_d^{\mathsf{XY}}(H(x) \times Y)(1+\epsilon)} \right| \tag{117}$$

$$< 2\epsilon \tag{118}$$

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

$$|R_d^A(x) - R_{d'}^A(x)| \leq \inf_n \left| R_d^A(x) - R_d^{A,n}(x) + R_d^{A,n}(x) - R_{d'}^{A,n}(x) + R_{d'}^{A,n} - R_{d'}^A \right| \tag{119}$$

$$\leq \inf_n \left(|R_d^A(x) - R_d^{A,n}(x)| + |R_d^{A,n}(x) - R_{d'}^{A,n}(x)| + |R_{d'}^{A,n} - R_{d'}^A| \right) \tag{120}$$

$$\leq 2\epsilon \tag{121}$$

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 $A \in \mathcal{Y}$

$$S^{A,n|\mathsf{XD}} := (x,d) \mapsto \mathbb{1}_{X \setminus O}(x) \sum_i^n \mathbb{1}_{\left[\frac{di}{n}, \frac{di+1}{n}\right]}(d) R_{\frac{di}{i}}^A(x) + \mathbb{1}_O(x)\delta_{y_0}(A) \tag{122}$$

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:

$$S_d^{A|\mathsf{XD}}(x) := \mathbb{1}_{X \setminus O}(x) R_d^A(x) + \mathbb{1}_O(x)\delta_{y_0}(A) \tag{123}$$

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

$$\int_B S_d^{A|\mathsf{XD}}(x) d\mathbb{K}_d^{\mathsf{X}}(x) = \int_B (\mathbb{1}_{X \setminus O}(x) R_d^A(x) + \mathbb{1}_O(x)\delta_{y_0}(A)) d\mathbb{K}_d^{\mathsf{X}} \tag{124}$$

$$= \int_B R_d^A(x) d\mathbb{K}_d^{\mathsf{X}}(x) \tag{125}$$

$$= \mathbb{K}_d^{\mathsf{XY}}(B \times A) \tag{126}$$

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 0.1.24 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 0.1.25 (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 $d \in D$, $B \times A \in \mathcal{X} \otimes \mathcal{Y}$*

$$\mathbb{K}_d^{\mathsf{XY}}(B \times A) = \sum_{J_i \in \mathcal{J}} \mathbb{1}_{J_i}(d) \mathbb{K}_d^{i,\mathsf{XY}}(B \times A) \tag{127}$$

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 0.1.24 we have for each i a disintegration $\mathbb{K}^{i,\mathsf{Y}|\mathsf{XD}}$. Define

$$T_{x,d}^{\mathsf{Y}|\mathsf{XD}}(A) := \sum_{J_i \in \mathcal{J}} \mathbb{1}_{J_i}(d) \mathbb{K}_{x,d}^{i,\mathsf{Y}|\mathsf{XD}}(A) \tag{128}$$

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 0.1.26. *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 0.1.27. *$Q_d^{A,n}$ as define in Theorem 0.1.24 is uniformly integrable.*

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

□

Theorem 0.1.28 (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 0.1.29 (Relative probability space)**.**

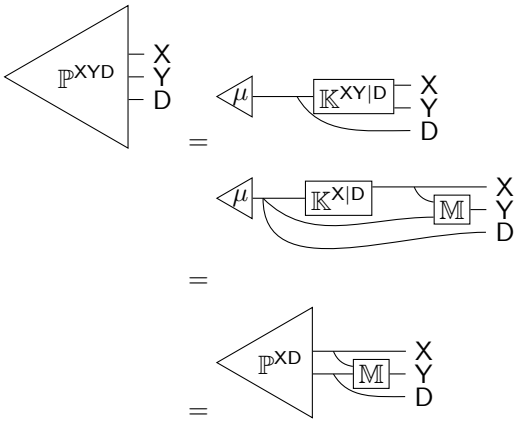
better name

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 X.

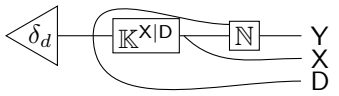
Lemma 0.1.30 (Agreement of disintegrations)**.** *Given a Markov kernel space* $(\mathbb{K}, D, \Omega)$*, any relative probability space* $(\mu\mathbb{K}, \Omega \times D)$ *and any random variables* $\mathsf{X} : \Omega \times D \to X$*,* $\mathsf{Y} : \Omega \times D \to Y$*,* $\mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}} = (\mu\mathbb{K})^{\{\mathsf{Y}|\mathsf{XD}\}}$ *(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



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 $d \in D$



$$\mathbb{Q} := \tag{132}$$

$$= \mathbb{K}_d^{\mathsf{XYD}|\mathsf{D}} \tag{133}$$

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

$$\mathbb{Q}(A \times B \times \{d\}) = \int_{X^2} \int_X \int_{D^3} \mathbb{N}_{d'',x'}(A)\mathbf{Id}_{x''}(B)\mathbf{Id}_{d'''}(\{d\}) d\curlyvee_d(d', d'', d''') d\mathbb{K}_{d'}^{\mathsf{X}|\mathsf{D}}(x) d\curlyvee_x(x', x'') \tag{134}$$

$$= \delta_d(\{d\}) \int_X \mathbb{N}_{d,x}(A)\delta_x(B) d\mathbb{K}_d^{\mathsf{X}|\mathsf{D}}(x) \tag{135}$$

$$= \frac{1}{\mu(\{d\})} \int_{\{d\}} d\mu(d') \int_X \mathbb{N}_{d,x}(A)\delta_x(B) d\mathbb{K}_d^{\mathsf{X}|\mathsf{D}}(x) \tag{136}$$

$$= \frac{1}{\mu(\{d\})} \int_D \int_X \mathbb{N}_{d,x}(A)\delta_{d'}(\{d\})\delta_x(B) d\mathbb{K}_d^{\mathsf{X}|\mathsf{D}}(a) d\mu(d') \tag{137}$$

$$= \frac{1}{\mu(\{d\})} \int_D \int_X \mathbb{N}_{d,x}(A)\delta_{d'}(\{d\})\delta_x(B) d\mathbb{K}_{d'}^{\mathsf{X}|\mathsf{D}}(a) d\mu(d') \tag{138}$$

$$= \frac{1}{\mu(\{d\})} \mathbb{P}^{\mathsf{XYD}}(A \times B \times \{d\}) \tag{139}$$

$$= \frac{1}{\mu(\{d\})} \int_D \mathbb{K}_{d'}^{\mathsf{XYD}|\mathsf{D}}(A \times B \times \{d\}) d\mu(d') \tag{140}$$

$$= \frac{1}{\mu(\{d\})} \int_D \mathbb{K}_{d'}\mathsf{XY}|\mathsf{D}(A \times B)\delta_{d'}(\{d\}) d\mu(d') \tag{141}$$

$$= \mathbb{K}_d^{\mathsf{XY}|\mathsf{D}}(A \times B) \tag{142}$$

$$= \mathbb{K}_d^{\mathsf{XY}|\mathsf{D}}(A \times B)\delta_d(\{d\}) \tag{143}$$

$$= \int_D \mathbb{K}_{d'}^{\mathsf{XY}}(A \times B)\delta_{d''}(\{d\}) d\curlyvee_d(d', d'') \tag{144}$$

$$= \mathbb{K}_d^{\mathsf{XYD}|\mathsf{D}}(A \times B \times \{d\}) \tag{145}$$

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 D is undefined in the kernel space $(\mathbb{K}, D, \Omega)$.

Conditional Independence

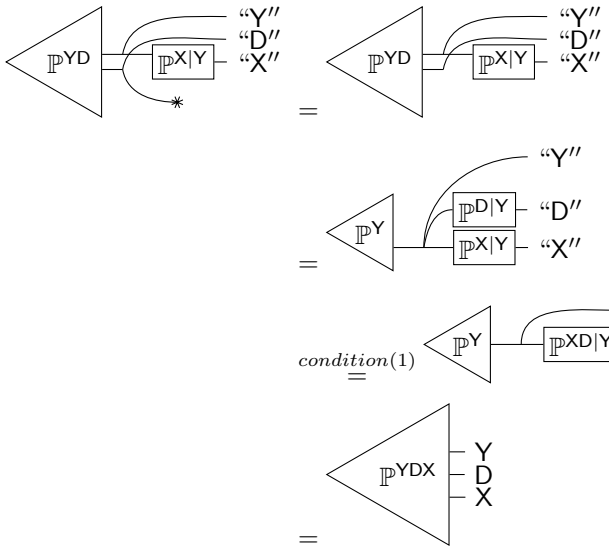
Definition 0.1.31 (Kernels constant in an argument)**.** Given a kernel $(\mathbb{K}, D, \Omega)$ and random variables Y and X, we say a verstion of the disintegration $\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}$ is constant in D if for all $x \in X$, $d, d' \in D$, $\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}_{(x,d)} = \mathbb{K}^{\mathsf{Y}|\mathsf{XD}}_{(x,d')}$.

Definition 0.1.32 (Domain Conditional Independence)**.** Given a kernel space $(\mathbb{K}, D, \Omega)$, relative probability space $(\mathbb{P}, \Omega \times D)$, variables X,Y and domain variable D, X is *conditionally independent* of D given 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 D

Theorem 0.1.33 (Definitions are equivalent)**.** *(1) $\implies$ (2): By Lemma 0.1.30, $\mathbb{P}^{\{\mathsf{Y}|\mathsf{XD}\}} = \mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$. Thus it is sufficient to show that $\mathbb{P}^{\mathsf{X}|\mathsf{Y}} \otimes *$ is a version of $\mathbb{P}^{\{\mathsf{X}|\mathsf{YD}\}}$.*



(146)

(147)

(148)

(149)

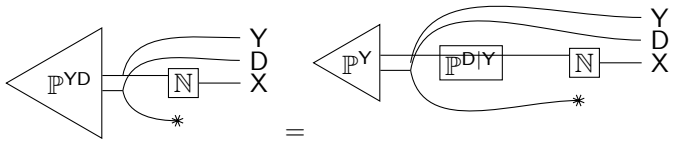
(2) $\implies$ (3)

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

(3) $\implies$ (1)

By lemma 0.1.30, there also exists a version of $\mathbb{P}^{\{\mathsf{X}|\mathsf{YD}\}}$ constant in 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 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 0.1.21, we have



(150)

Definition 0.1.34 (Conditional probability existence)**.** Given a kernel space $(\mathbb{K}, D, \Omega)$ and random variables X, 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 0.1.35 (Conditional Independence)**.** Given a kernel space $(\mathbb{K}, D, \Omega)$, some relative probability space $(\mathbb{P}, \Omega \times D)$, variables X,Y and Z, X is *conditionally independent* of Z given 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 Z

Lemma 0.1.36 (Diagrammatic consequences of labels)**.** *In general, diagram labels are "well behaved" with regard to the application of any of the special Markov kernels: identities 17, swaps 28, discards 34 and copies 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 X, Y, Z taking values in X, Y, Z respectively. Sat : *indicates that a labeled diagram satisfies definitions 0.1.16 and 0.1.19 with respect to $(\mathcal{K}, D, \Omega)$ and X, Y, Z. The following always holds:*

$$\text{Sat} : \mathsf{X} - \mathsf{X} \tag{151}$$

and the following implications hold:

$$\text{Sat} : \mathsf{Z} \multimap \boxed{\mathbb{K}} \multimap \begin{matrix}\mathsf{X}\\ \mathsf{Y}\end{matrix} \implies \text{Sat} : \mathsf{Z} \multimap \boxed{\mathbb{K}} \multimap \mathsf{X} \multimap * \tag{152}$$

$$\text{Sat} : \mathsf{Z} \multimap \boxed{\mathbb{K}} \multimap \begin{matrix}\mathsf{X}\\ \mathsf{Y}\end{matrix} \implies \text{Sat} : \mathsf{Z} \multimap \boxed{\mathbb{K}} \multimap \begin{matrix}\mathsf{Y}\\ \mathsf{X}\end{matrix} \tag{153}$$

$$\text{Sat} : \mathsf{Z} \multimap \boxed{\mathbb{L}} \multimap \mathsf{X} \implies \text{Sat} : \mathsf{Z} \multimap \boxed{\mathbb{L}} \multimap \begin{matrix}\mathsf{X}\\ \mathsf{X}\end{matrix} \tag{154}$$

$$\text{Sat} : \mathsf{Z} \multimap \boxed{\mathbb{K}} \multimap \mathsf{Y} \implies \text{Sat} : \begin{matrix}\mathsf{Z} \multimap \\ \mathsf{Z} \multimap \boxed{\mathbb{K}} \multimap \mathsf{Y}\end{matrix} \tag{155}$$

Proof.

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

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

$$\mathbb{P}_{\mathsf{ZY}} = \begin{matrix}\mathbb{P}_{\mathsf{Z}} \multimap \mathsf{Z}\\ \qquad\quad \boxed{\mathbb{K}} \multimap \mathsf{Y}\end{matrix} \tag{156}$$

$$\mathbb{P}_{\mathsf{ZZY}} = \begin{matrix}\mathbb{P}_{\mathsf{Z}} \multimap \mathsf{Z}\\ \qquad\quad \mathsf{Z}\\ \qquad\quad \boxed{\mathbb{K}} \multimap \mathsf{Y}\end{matrix} \tag{157}$$

$$= \begin{matrix}\mathbb{P}_{\mathsf{Z}} \multimap \mathsf{Z}\\ \qquad\quad \mathsf{Z}\\ \qquad\quad \boxed{\mathbb{K}} \multimap \mathsf{Y}\end{matrix} \tag{158}$$

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

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

$$\text{Sat} : \mathsf{Z} \multimap \boxed{\mathbb{K}} \multimap \mathsf{Y} \ , \ \mathsf{Y} \multimap \boxed{\mathbb{L}} \multimap \mathsf{X} \implies \text{Sat} : \mathsf{Z} \multimap \boxed{\mathbb{K}} \multimap \boxed{\mathbb{L}} \multimap \mathsf{X} \tag{159}$$

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

Chapter 1

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 β 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" θ of voters for each candidate and, under some unreasonably strong sampling assumptions, and the results of the survey for each θ can

be described by $\prod_N \mathbb{P}_\theta^{\mathsf{X}}$ where N is the number of voters surveyed and 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 α, which β 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. Their general types are the same, and the only difference is in the interpretation of the sets H and D. 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 so-called 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.

Decision problems involving often involve some data X is observed, then a choice is made, then the consequences Y are observed. In such a model, the observed data 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.

In this chapter I will introduce two-player statistical models "looking backwards" towards key results in the foundations of ordinary statistics, with a particular focus on exchangeability-like assumptions and statistical decision theory. In the following chapters, results obtained here will be applied "looking forwards" problems of causal inference.

1.1 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 X, 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 0.1.24).

Definition 1.1.1 (Two player statistical model)**.** A *two-player statistical model* $(\mathbb{T}, \mathsf{H}, \mathsf{D}, \mathsf{O})$ is a Markov kernel $\mathbb{T} : H \times D \to \Delta(\mathcal{E})$ such that, for any random variables $\mathsf{X} : H \times D \times E \to X$ and $\mathsf{Y} : H \times D \times E \to Y$, 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* $\mathsf{H} : H \times D \times E \to H$ given by $(h, d, e) \mapsto h$ (forgetting

the choice and outcome) and the *choice* $\mathsf{D} : H \times D \times E \to D$ given by $(h, d, e) \mapsto d$ (forgetting the hypothesis and outcome) and the *outcome* $\mathsf{O} : H \times D \times E \to E$ given by $(h, d, e) \mapsto e$ (forgetting the choice and hypothesis).

Definition 1.1.2 (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* $\mathsf{X} : H \times D \times E \to X$ and the *consequence* $\mathsf{Y} : H \times D \times E \to Y$ 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 O when specifying a see-do model.

1.1.1 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 1.1.3 (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.

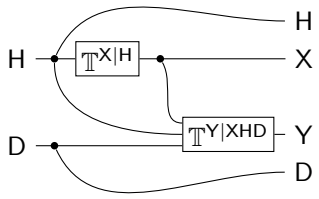
Decomposable see-do models can be represented as a pair $(\mathbb{B}, \mathbb{C})$ where $\mathbb{B}$ is a one-player statistical model we call the *observation model* and $\mathbb{C}$ is a two-player statistical model we call the *consequence model* (Corollary 1.1.5. Most models in the causal inference literature are decomposable – if the observed data can tell us nothing useful beyond the distribution of observations, then we have a decomposable model.

Theorem 1.1.4 (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

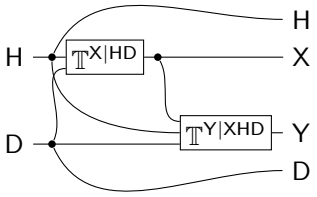
$\mathbb{T} =$



(1.1)

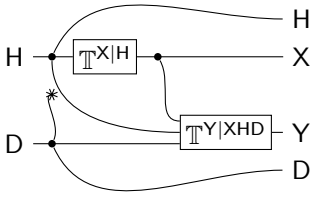
Maybe moves proofs out of main text

Proof. By 0.1.1,

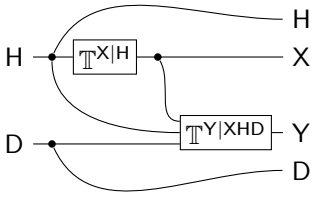


$$\mathbb{T} = \tag{1.2}$$

By the assumption $X \perp\!\!\!\perp_{\mathbb{T}} D|H$ and version 2 of conditional independence from Theorem 0.1.33,



$$\mathbb{T} = \tag{1.3}$$



$$= \tag{1.4}$$

□

Corollary 1.1.5. *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 $\mathbb{T}^{X|H}$
- *The* consequence model $\mathbb{T}^{Y|HD}$

Proof. Because $\mathbb{T}$ is decomposable, $\mathbb{T}^{Y|XHD} = *_X \otimes \mathbb{T}^{Y|HD}$. Then by theorem 0.1.1 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 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 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 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".

Exchangeability

Thus far, we haven't dwelt on what it means for a probability measure to "describe" or "represent" something. 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
2. Forecasts of observations that will take place in the future

Taking the first view, one can view each hypotheses in a statistical model as representing the proposition that the system will tend to produce long-run sequences of observations favoured by the associated probability measure. 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 so forth. On the other hand, if we view a sequence of random variables as a sequence of forecasts it is not immediately obvious that we need such hypotheses. If $\mathsf{X}_1, \mathsf{X}_2, \mathsf{X}_3, ...$ are rolls of a possibly loaded die, then it seems reasonable on observing a large number of sixes that sixes are more likely in the future. Unlike hypotheses a) and b), forecasts of the outcome do not assign a stable value to the proportion of sixes produced by the die.

The first view is usually called *frequentist* and the second view is usually called *Bayesian*. Discussions of frequentism and Bayesianism often carry overtones of whether we should consider probabilities to represent long run frequencies or forecasts, and this issue is much debated Hájek (2019).

When probability measures are used to represent convergence of relative frequencies, one player statistical models $(\mathbb{T}, \mathsf{H}, \mathsf{O})$ are often used to describe the system of interest, which include a hypothesis H as a basic element. When probability measures are used to represent forecasts of future observations, we typically only consider a single probability measure $\mathbb{F} \in \Delta(\mathcal{E})$ along with an outcome variable O. These are not hard requirements of either view – someone modelling frequencies might have reason to model a system with only one hypothesis, and a forecaster might want to consider a set of forecasts generated

by a number of hypotheses. Nonetheless, the two purposes typically differ on whether hypotheses are a basic element of the system model.

The analogy with two player statistical models is that we can consider a two player model $(\mathbb{T}, \mathsf{H}, \mathsf{D}, \mathsf{O})$ to describe the long run frequencies obtained by repeated experiments on a system, or to describe a *forecast* of what is likely to happen given a particular choice we can use a map $\mathbb{F} : D \to \Delta(\mathcal{E})$, the observation D and the outcome O. We will call such objects *forecasts*, and *see-do forecasts* the special subset of forecasts that feature observations that are independent of choices.

Definition 1.1.6 (Forecasts, see-do forecast)**.** A *do forecast* $(\mathbb{F}, \mathsf{D}, \mathsf{O})$ is a Markov kernel $\mathbb{F} : D \to \Delta(\mathcal{E})$ for some set of choices D and outcomes E. The choice variable $\mathsf{D} : D \times E \to D$ is the map $(d, e) \mapsto d$ that forgest the outcome and and the outcome variable $\mathsf{O} : D \times E \to E$ is the map $(d, e) \to e$ that forgets the choice.

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

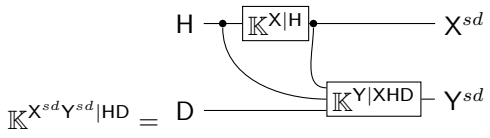
A *forecast* $(\mathbb{F}, \mathsf{O})$ is a probability measure $\mathbb{F} \in \Delta(\mathcal{E})$ and an outcome variable $\mathsf{O} : E \to O$.

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

Theorem 1.1.7. *The product* $\mathbb{L} := (\mu \otimes \mathrm{Id}_D)\mathbb{K}$ *of a* prior $\mu \in \Delta(\mathcal{H})$ *and a two player statistical model* $\mathbb{K} : H \times D \to \Delta(\mathcal{E})$ *is a forecast* $D \to \Delta(\mathcal{E})$ *with* $\mathsf{D}^f : (d, e) \mapsto d$ *and* $\mathsf{E}^f : (d, e) \mapsto e$*. If* $\mathbb{K}$ *is a see-do model with respect to observations* X^{sd}, Y^{sd} *then there exist* X^f, Y^f *such that* $\mathbb{L}$ *is a see-do forecast.*

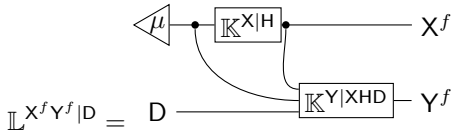
Proof. The first part is trivial: $(\mu \otimes \mathrm{Id}_D)\mathbb{K}$ is a Markov kernel $D \to \Delta(\mathcal{E})$ by construction.

For the second part $\mathsf{X}^f \perp\!\!\!\perp \mathsf{D}$ is required for some X^f.By Theorem 0.1.21 we have



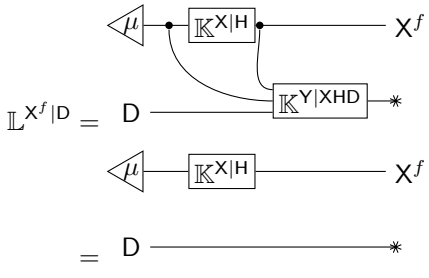
(1.5)

Define X^f and Y^f such that



(1.6)

Clearly $\mathsf{O}^f = \mathsf{X}^f \underline{\otimes} \mathsf{Y}^f$.
Then



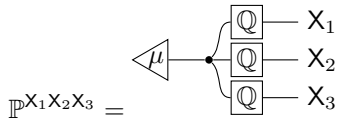
(1.7)

(1.8)

And so $\mathsf{X}^f \perp\!\!\!\perp_{\mathbb{M}} \mathsf{D}$. □

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 0.1.3, 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, a statistical model with a single hypothesis is a very unusual type of statistical model.

de Finetti (1992) has shown how more typical statistical models can be recovered from certain types of forecast. 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 $\mathbb{Q} : H \to \Delta(\mathcal{E})$ and a *prior* $\mu \in \Delta(\mathcal{H})$ such that



(1.9)

Defining the hypothesis $\mathsf{H} : E \mapsto H$ 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.

Exchangeability-like assumptions have a number of interesting applications to see-do models. In addition to providing a means of obtaining see-do models from see-do forecasts analogous to the theorm of de Finetti (1992) mentioned above, we have a number of additional results:

- Exchangeability of observations implies decomposability

- We show that the existence of counterfactual random variables is equivalent to *functional exchangeability* along with *deterministic reproducibility* and *non-interference*
- *Imitable* see-do models, a special class of doubly exchangeable models, play a key role in identification of causal effects

Not happy with the previous list

Definition 1.1.8 (Permutations and swaps)**.** A *finite permutation* ρ' 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 $\rho : E \to \Delta(\mathcal{E})$ associated with a finite permutation ρ' is a Markov kernel associated with the fuction ρ^* which has the property $\mathsf{X}_i \circ \rho^* = \mathsf{X}_{\rho'(i)}$ for all $i \in B$.

If $E = X_0^{|B|}$ and $\mathsf{X}_i : E \to X_1$ projects the i-th element of the space $(x_1, ..., x_i, ...) \mapsto x_i$, then for some ρ' the associated swap ρ is the Markov kernel associated with the fuction $\rho^* : (x_1, ..., x_i, ...) \mapsto (x_{\rho'(1)}, ..., x_{\rho'(i)}, ...)$.

A consequence of this is

$$\rho \underline{\otimes}_{i\in B} \mathbb{F}_{\mathsf{X}_i} = \underline{\otimes}_{i\in B} \rho \mathbb{F}_{\mathsf{X}_i} \tag{1.10}$$

$$= \underline{\otimes}_{i\in B} \mathbb{F}_{\mathsf{X}_{\rho'(i)}} \tag{1.11}$$

Where line 1.10 follows from the fact that deterministic kernels commute with the split map (159), and line 1.11 follows from the fact that for two functional kernels

$$(\mathbb{F}_f \mathbb{F}_g)_x(A) = \int_X (\mathbb{F}_g)_y(A) d(\mathbb{F}_f)_x(y) \tag{1.12}$$

$$= \int_X \delta_{g(y)}(A) d\delta_{f(x)}(y) \tag{1.13}$$

$$= \delta_{g(f(x))}(A) \tag{1.14}$$

$$= (\mathbb{F}_{g\circ f})_x(A) \tag{1.15}$$

lemmafy, move to chapter 2

Definition 1.1.9 (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 ρ is the swap associated with ρ' and $\{\mathsf{X}_i | i \in B\}$, then $(\mathbb{T}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ is *exchangeable* with respect to $\{\mathsf{X}_i | i \in A\}$.

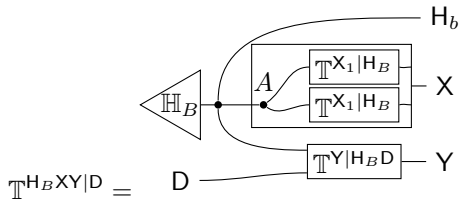
If A is an infinite set then $\mathbb{T}$ is *infinitely exchangeable*, and if $\mathbb{T} = \mathbb{S}(\mathrm{Id}_X \otimes * \otimes \mathrm{Id}_Y)$ for some infinitely exchangeable $(\mathbb{S}, \mathsf{D}, \mathsf{X}', \mathsf{Y}')$, then $\mathbb{T}$ is infinitely exchangeably extendable.

Note that $\mathbb{T}\rho^{\mathsf{X}_i|\mathsf{D}} = \mathbb{T}^{\mathsf{X}_{\rho'(i)}|\mathsf{D}}$.

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

Theorem 1.1.10 (Representation of infinitely exchangeably extendable see-do forecasts)**.** *Given a see-do forecast* $(\mathbb{T}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ *where* $\mathsf{X} = \underline{\otimes}_{i\in A}\mathsf{X}_i$ *for some* $\{\mathsf{X}_i | i \in A \subseteq \mathbb{N}\}$ *and* $X \times Y$ *is standard measurable, the following statements are equivalent:*

1. $(\mathbb{T}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ *is infinitely exchangeably extendable with respect to* $\{\mathsf{X}_i | i \in A\}$
2. *There exists a function* $f : X \to H_B$ *such that, defining* $\mathsf{H}_B : f \circ \mathsf{X}$,



(1.16)

3. *There exists a function* $f : X \to H_B$ *such that, defining* $\mathsf{H}_B : f \circ \mathsf{X}$*, for any* $d \in D$*,* $L \in \mathcal{H}_B$ $(\times_{i\in A} J_i) \in \mathcal{X}$*,* $K \in \mathcal{Y}$

$$\mathbb{T}_d^{\mathsf{H}_B\mathsf{X}\mathsf{Y}|\mathsf{D}}(L \times (\times_{i\in A} J_i) \times K) = \int_L \prod_{i\in A} \mathbb{T}_h^{\mathsf{X}_1|\mathsf{H}_B}(J_i)\mathbb{T}_{h,d}^{\mathsf{Y}|\mathsf{H}_B\mathsf{D}}(K)d\mathbb{T}^{\mathsf{H}_B}(h) \tag{1.17}$$

4. *There exists a function* $f : X \to H_B$ *such that, defining* $\mathsf{H}_B : f \circ \mathsf{X}$
   - $\mathsf{X}_i \perp\!\!\!\perp \mathsf{X}_{A\setminus\{i\}} | \mathsf{H}_B$ *for all* $i \in A$
   - $\mathbb{T}^{\mathsf{X}_i|\mathsf{H}_B} = \mathbb{T}^{\mathsf{X}_j|\mathsf{H}_B}$ *for all* $i, j \in A$
   - $\mathsf{Y} \perp\!\!\!\perp \mathsf{X} | \mathsf{H}_B \underline{\otimes} \mathsf{D}$

Proof. (2) $\iff$ (3) follows from the fact that (3) + the associated quantifications is the integral representation of the string diagram in (2).

We will show (3) $\implies$ (1), (1) $\implies$ (4) and (4) $\implies$ (3).

(3) $\implies$ (1):

we have

$$\mathbb{T}_d((\times_{i\in A} J_i) \times K) = \int_{H_B} \prod_{i\in A} \mathbb{T}_h^{\mathsf{X}_1|\mathsf{H}_B}(J_i)\mathbb{T}_{h,d}^{\mathsf{Y}|\mathsf{H}_B\mathsf{D}}(K)d\mathbb{T}^{\mathsf{H}_B}(h) \tag{1.18}$$

For any $(\times_{i\in\mathbb{N}} J_i) \in {\mathcal{X}_\infty}^{\mathbb{N}}$, define

$$\mathbb{U}_d((\times_{i\in\mathbb{N}}J_i)\times K)) = \int_{H_B}\prod_{i\in C}\mathbb{T}_h^{\mathsf{X}_1|\mathsf{H}_B}(J_i)\mathbb{T}_{h,d}^{\mathsf{Y}|\mathsf{H}_B\mathsf{D}}(K)d\mathbb{T}^{\mathsf{H}_B}(h) \tag{1.19}$$

We have

$$\mathbb{U}_d(\mathrm{Id}_X\otimes *\otimes \mathrm{Id}_Y)(L\times(\times_{i\in A}J_i)\times K) = \int_L\prod_{i\in A}\mathbb{T}_h^{\mathsf{X}_1|\mathsf{H}_B}(J_i)\prod_{i\in\mathbb{N}\setminus A}(1)\mathbb{T}_{h,d}^{\mathsf{Y}|\mathsf{H}_B\mathsf{D}}(K)d\mathbb{T}^{\mathsf{H}_B}(h) \tag{1.20}$$

$$= \mathbb{T}_d^{\mathsf{H}_B\mathsf{X}\mathsf{Y}|\mathsf{D}}(L\times(\times_{i\in A}J_i)\times K) \tag{1.21}$$

Define $\{\mathsf{X}_i'|i\in\mathbb{N}\}$ where $\mathsf{X}_i : H_B\times X_1^{\mathbb{N}}\times Y\to X_1$ takes $(h,x_1,...,x_i,...,y)\mapsto x_i$ for all i, and similarly $\mathsf{Y}' : H_B\times X_1^{\mathbb{N}}\times Y\to Y$ takes $(h,x_1,...,y)\mapsto y$. Given any permutation $\rho':\mathbb{N}\to\mathbb{N}$

$$\mathbb{U}_d\rho(H_b\times(\times_{i\in\mathbb{N}}J_i)\times K) = \mathbb{U}_d(H_b\times(\times_{i\in\mathbb{N}}J_{\rho'(i)})\times K) \tag{1.22}$$

$$= \int_{H_B}\prod_{i\in\mathbb{N}}\mathbb{T}_h^{\mathsf{X}_1|\mathsf{H}_B}(J_{\rho'(i)})\mathbb{T}_{h,d}^{\mathsf{Y}|\mathsf{H}_B\mathsf{D}}(K)d\mathbb{T}^{\mathsf{H}_B}(h) \tag{1.23}$$

$$= \mathbb{U}_d(H_b\times(\times_{i\in\mathbb{N}}J_i)\times K) \tag{1.24}$$

Which shows $\mathbb{U}_d$ is infinitely exchangeable, completing the proof that (3) $\implies$ (1)

(1) $\implies$ (4):

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 $\mathcal{Z}_q^n=\sigma(\mathsf{Z}_q^n)$. For any swap $\rho : D\times X\times Y\to D\times X\times Y$ that swaps X_is according to some finite permuation $\rho':\mathbb{N}\to\mathbb{N}$ X_is $(d,x_1,x_2,...,y)\mapsto(d,x_{\rho'(1)},x_{\rho'(d)},...,y)$, we have $\mathsf{Z}_q^n\circ\rho=\mathsf{Z}_q^n$. Thus $(\mathsf{Z}_q^n)^{-1}(A)=(\mathsf{Z}_q^n\circ\rho^{-1})^{-1}(A)=\rho(\mathsf{Z}_q^n)^{-1}(A)$.

Note that $\mathcal{Z}_q^1\supset\mathcal{X}_q^2\supset,...\mathcal{Z}_q:=\int_i^{\mathbb{N}}\mathcal{Z}_q^i$.

Let $\rho_{ij}':\mathbb{N}\to\mathbb{N}$ be the permutation that swaps indices i and $j$for any $j\in[n]$, and $\rho_{ij} : D\times X\times Y\to D\times X\times Y$ the swap kernel associated with ρ_{ij}' and $\{\mathsf{X}_i|i\in\mathbb{N}\}$, and ρ_{ij}^* the function associated with ρ_{ij}. For any m, n,A$\in\mathcal{Z}_1^n$, $d\in D$:

$$\int_A \mathsf{Z}_q^m(\omega)d\mathbb{T}_d(\omega) = \int_A \frac{1}{m}\sum_i^m \mathbb{1}_{[0,q]}(\mathsf{X}_i(\omega))d\mathbb{T}_d(\omega) \tag{1.25}$$

$$= \frac{1}{m}\sum_i^m \int_{\rho_{ij}^{*\,-1}\rho_{ij}^*(A)} \mathbb{1}_{[0,q]}(\mathsf{X}_i(\omega))d\mathbb{T}_d\rho_{ij}(\omega) \tag{1.26}$$

$$= \frac{1}{m}\sum_i^m \int_{\rho_{ij}^*(A)} \mathbb{1}_{[0,q]}(\mathsf{X}_i \circ \rho_{ij}(\omega))d\mathbb{T}_d(\omega) \tag{1.27}$$

$$= \frac{1}{m}\sum_i^m \int_A \mathbb{1}_{[0,q]}(\mathsf{X}_j(\omega))d\mathbb{T}_d(\omega) \tag{1.28}$$

$$= \int_A \mathbb{1}_{[0,q]}(\mathsf{X}_j(\omega))d\mathbb{T}_d(\omega) \tag{1.29}$$

Where line 1.26 follows from exchangeability of $\mathbb{T}$ and invertibility of ρ_{ij}. the fact that ρline 1.27 follows from the fact that $\mathbb{T}_d\rho_{ij}$ is the pushforward measure of $\mathbb{T}_d$ with respect to ρ_{ij}^* and 1.28 uses the fact that $\rho(A) = A$ for all $A \in \mathsf{Z}_q^n$ and all permutations ρ.

From Equation 1.29, we have

$$\int_A \mathsf{Z}_q^m(\omega)d\mathbb{T}_d(\omega) = \int_A \tag{1.30}$$

Because $\mathsf{X}_{[n]}^{-1}(A) = \mathsf{X}_{[n+1]}^{-1}(A \times X_1)$ and $\{\mathsf{X}_{[n]}^{-1}(A)|A \in \mathbb{N}\}$ generates $\mathcal{H}_n$, we have for all $B \in \mathcal{H}_n$

$$\int_B \mathsf{Z}^{n+1}(\omega)d\mathbb{T}_d(\omega) = \int_B \mathsf{Z}^n(\omega)d\mathbb{T}_d(\omega) \tag{1.31}$$

Because Z^n is $\mathcal{H}^n$-measurable, Eq. 1.31 shows that Z^n is a version of $\mathbb{E}[\mathsf{Z}^{n+1}|\mathcal{H}^n]$ and hence the sequence $\{\mathsf{Z}_n\}_{\mathbb{N}}$ is a martingale.

Furthermore, for all $n \in \mathbb{N}$, $|\mathsf{Z}^n| \leq 1$ so the sequence is also uniformly integrable. Thus it goes almost surely to a limit Z and for all $B \in \mathcal{B}$

$$\lim_{n\to\infty} \int_B \mathsf{Z}^n(\omega)d\mathbb{T}_d(\omega) = \int_B \mathsf{Z}(\omega)d\mathbb{T}_d(\omega) \tag{1.32}$$

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

$$\int_B \mathbb{1}_{[0,q]}(\mathsf{X}_j(\omega))d\mathbb{T}_d(\omega) = \int_B \mathsf{Z}^n(\omega)d\mathbb{T}_d(\omega) \tag{1.33}$$

it follows that for all $B \in \mathcal{H}$

$$\int_B \mathsf{Z}(\omega)d\mathbb{T}_d(\omega) = \int_B \mathbb{1}_{[0,q]}(\mathsf{X}_j(\omega))d\mathbb{T}_d(\omega) \tag{1.34}$$

[Çinlar (2011) Thm 4.7.]. Thus $\mathsf{Z} = \mathbb{E}[\mathbb{1}_{[0,q]}(\mathsf{X}_j|\mathcal{H}]$ for all $j \in \mathbb{N}$. For each □

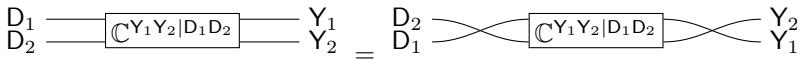
$$\mathsf{H}_B^{(i)} = \underline{\otimes}_{q\in Q}\mathbb{1}_{[0,q]} \circ \mathsf{X}_i \tag{1.35}$$

$$\mathsf{H}_B^n = \underline{\otimes}_{i\in[n]}\mathsf{H}_B^{(i)} \tag{1.36}$$

Definition 1.1.11 (Functional Exchangeability)**.** Given a two player statistical model $\mathbb{K} : D \times H \to \Delta(\mathcal{E})$, a *choice variable* D_i is a variable such that $\mathsf{D}_i \perp\!\!\!\perp_{\mathbb{K}} \mathsf{O}|\mathsf{D}$ a finite sequence of random variables $\mathsf{Y}_1, ..., \mathsf{Y}_n : D \times E \to Y$ and a sequence of choice variables $\mathsf{D}_1, ..., \mathsf{D}_n : D \to D_0$ on kernel space $(\mathbb{C}, D, \Omega)$ along with is *functionally exchangeable* if for all permutations $\sigma : [n] \to [n]$, $\mathbb{C}^{\mathsf{Y}_1,...,\mathsf{Y}_n|\mathsf{D}_1,...,\mathsf{D}_n} = \mathbb{C}^{\mathsf{Y}_{\sigma(1)},...,\mathsf{Y}_{\sigma(n)}|\mathsf{D}_{\sigma(1)},...,\mathsf{D}_{\sigma(n)}}$.

define choice variables; independent of Y conditional on D

Graphically, functional exchangeability of $\mathsf{Y}_1, \mathsf{Y}_2$ and $\mathsf{D}_1, \mathsf{D}_2$ implies

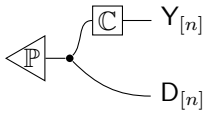


(1.37)

Include a lemma about swap maps and variable permutations

Lemma 1.1.12 (Functionally exchangeable sequences with exchangeable choices induce exchangeable sequences)**.** *Given functionally exchangeable sequences* $\mathsf{Y}_{[n]}$ *and* $\mathsf{D}_{[n]}$ *on* $(\mathbb{C}, D_0^n, Y_0^n)$ *with product* σ*-algebra, along with an exchangable measure* $\mathbb{P}^{\mathsf{D}_{[n]}}$*, define* $\mathbb{P}'$ *as follows:*

$\mathbb{P}' =$



(1.38)

Then sequence $\underline{\otimes}_{i\in[n]}(\mathsf{Y}_i \otimes \mathsf{D}_i)$ *(given by the obvious projection maps) on the probability space* $(\mathbb{P}', Y_0^n \times D_0^n, \mathcal{Y}_0^n \otimes \mathcal{D}_0^n)$ *is exchangeable.*

Proof. For $i \in [n]$, $A_i \in \mathcal{Y}$, $B_i \in \mathcal{D}$ and arbitrary permutation σ we have

$$\mathbb{P}'^{\mathsf{X}_1\mathsf{D}_1,\ldots,\mathsf{X}_n\mathsf{D}_n}(\prod_{i\in[n]} A_i \times B_i) = \int_{\prod_{i\in[n]} B_i} \mathbb{P}'^{\mathsf{X}_{[n]}|\mathsf{D}_{[n]}}_{d_{[n]}}(\prod_{i\in[n]} A_i) d\mathbb{P}'^{\mathsf{D}_{[n]}}(d_{[n]}) \quad (1.39)$$

$$= \int_{\prod_{i\in[n]} B_i} \mathbb{C}^{\mathsf{X}_{[n]}|\mathsf{D}_{[n]}}_{d_{[n]}}(\prod_{i\in[n]} A_i) d\mathbb{P}^{\mathsf{D}_{[n]}}(d_{[n]}) \quad (1.40)$$

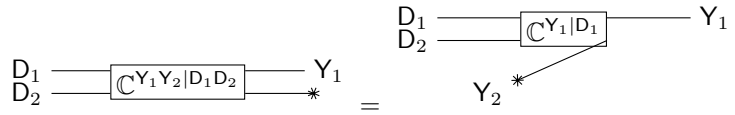
$$= \int_{\prod_{i\in[n]} B_i} \mathbb{C}^{\mathsf{X}_{\sigma([n])}|\mathsf{D}_{\sigma([n])}}_{d_{[n]}}(\prod_{i\in[n]} A_i) d\mathbb{P}^{\mathsf{D}_{\sigma([n])}}(d_{[n]}) \quad (1.41)$$

$$= \int_{\prod_{i\in[n]} B_i} \mathbb{P}'^{\mathsf{X}_{\sigma([n])}|\mathsf{D}_{\sigma([n])}}_{d_{[n]}}(\prod_{i\in[n]} A_i) d\mathbb{P}'^{\mathsf{D}_{\sigma([n])}} \quad (1.42)$$

$$= \mathbb{P}'^{\mathsf{X}_{\sigma(1)}\mathsf{D}_{\sigma(1)},\ldots,\mathsf{X}_{\sigma(n)}\mathsf{D}_{\sigma(n)}}(\prod_{i\in[n]} A_i \times B_i) \quad (1.43)$$

Where line 1.41 follows from exchangeability of $\mathbb{P}$ and functional exchangeability of $\mathbb{C}$ and lines 1.40 and 1.42 follow from the fact that for any invertible function f of $\mathsf{D}_{[n]}$ and random variable Y, $\mathbb{C}^{\mathsf{Y}|f(\mathsf{D}_{[n]})}$ is a version of $\mathbb{P}'^{\mathsf{Y}|f(\mathsf{D}_{[n]})}$ and $\mathbb{P}^{f(\mathsf{D}_{[n]})} = \mathbb{P}'^{\mathsf{Y}|f(\mathsf{D}_{[n]})}$. □

Definition 1.1.13 (Non-interfering)**.** A pair of sequences $\mathsf{Y}_{[n]}$ and $\mathsf{D}_{[n]}$ on $(\mathbb{C}, D, \Omega)$ is *noninterfering* if for all $U \subset [n]$, $\mathbb{C}^{\mathsf{Y}_U|\mathsf{D}_U}$ exists and $\mathbb{C}^{\mathsf{Y}_U|\mathsf{D}_{[n]}} = \mathbb{C}^{\mathsf{Y}_U|\mathsf{D}_U}$. Non-interference implies that discard maps can "fall through" kernels:



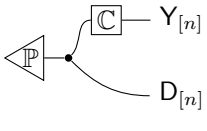
(1.44)

I think it is the case that functionally exchangeable + non-interfering implies infinitely functionally exchangeably extendable, but not proved yet

Theorem 1.1.14 (Representation of functionally exchangeable sequences)**.** *Let $(Y, \mathcal{Y})$ be a compact Hausdorff space with the Baire σ-algebra and $(D, \mathcal{D})$ a finite discrete space. Let $\mathbb{C}$ be a Markov kernel $D^{\mathbb{N}} \to \Delta(\mathcal{Y}^{\mathbb{N}})$ and $\mathsf{Y}_{\mathbb{N}}$, $\mathsf{D}_{\mathbb{N}}$ a pair of functionally exchangeable sequences. Define $(F, \mathcal{F})$ to be the set of all Markov kernels $D \to \Delta(\mathcal{Y})$ with $\mathcal{F}$ the coarsest σ-algebra for which all evaluation maps $\mathrm{ev}_{d,A} : F \to \mathbb{R}$ given by $\mathrm{ev}_{d,A} : \mathbb{H} \mapsto \mathbb{H}_d(A)$ are measurable. Then there exists a unique probability measure ν on $\Delta(\mathcal{F})$ such that for all $n \in \mathbb{N}$, $d \in D^n$, $C \in \mathcal{Y}^n$:*

$$\mathbb{C}_d(C) = \int_F \prod_{i\in[n]} \mathbb{H}_{\mathsf{D}_i(d)}(\mathsf{Y}_i(C)) d\nu(\mathbb{H}) \quad (1.45)$$

Proof. Let $\delta \in \Delta(\mathcal{D})$ be such that for all $n \in \mathbb{N}$, $\emptyset \neq B \in \mathcal{D}^n$ we have $\delta\curlyvee_n(B) > 0$. Such a measure exists by assumption on D. Define $delta_{\underline{n}} := \delta\curlyvee_n$. It is trivial to show that $\delta_{\underline{n}}$ is exchangeable. Define

$\mathbb{C}^n_\delta :=$ 

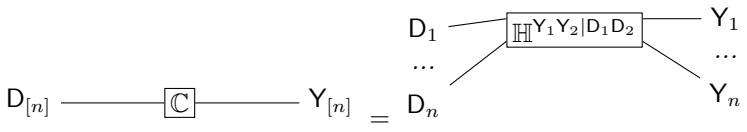
(1.46)

$:= \delta_{\underline{n}}\mathbb{C}\underline{\otimes}\mathrm{Id}_{D^n}$

. By Lemma 1.1.12, there is an exchangeable sequence $\underline{\otimes}_{i\in[n]}\mathsf{Y}'_i\underline{\otimes}\mathsf{D}'_i$ on the probability space $(\delta_{\underline{n}}\mathbb{C}\underline{\otimes}\mathrm{Id}_D, Y^n \times D^n, \mathcal{Y}^n \otimes \mathcal{D}^n)$.

□

Corollary 1.1.15. *Equivalently, for all $n \in \mathbb{N}$, let $\mathbb{G} : F \to$*



(1.47)

1.1.2 Causal questions and decision functions

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

1. Association: How are W and Z related? How would observing W change my beliefs about 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 X and I know which values of 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 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

Definition 1.1.16 (Consequence map)**.** Given a see-do model $(\mathbb{T}, \mathsf{H}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$, a *consequence map* is a map $\mathbb{C} : D \to \Delta(\mathcal{Y})$ where D is a choice set and Y is a consequence set.

The consequence model evaluated at any particular hypothesis $h \in H$, $\mathbb{T}^{\mathsf{Y}|\mathsf{XHD}}_{\cdot,h,\cdot}$ is a consequence map.

Not quite sure if this is the right place for the following definition

The independence of observations and choices is preserved when we take the product of a see-do model and a *prior* over hypotheses. Such a product produces a *Bayesian see-do model*:

Definition 1.1.17 (Bayesian See-Do Model)**.** A Bayesian See-Do Model $\langle \mathbb{U}, \mathsf{D}, \mathsf{X}, \mathsf{Y} \rangle$ is a Markov kernel space $(\mathbb{U}, D, X \times Y)$ with the property $\mathsf{X} \perp\!\!\!\perp_{\mathbb{U}} \mathsf{D}$, along with choices D, observations X and consequences Y, defined as before.

Theorem 1.1.18 (A see-do model with a prior is a Bayesian see-do model)**.** *The product of a see-do model $\mathbb{T}$ and a prior $\gamma \in \Delta(\mathcal{H})$*

$$\mathbb{U} := (\gamma \otimes \mathrm{Id}^D)\mathbb{T} \tag{1.48}$$

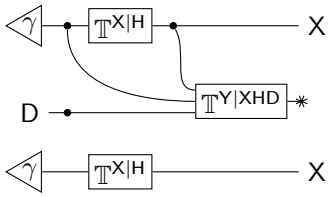
Is a Bayesian see-do model.

Proof. It nees to be shown that $\mathsf{X} \perp\!\!\!\perp_{\mathbb{U}} \mathsf{D}$.

By definition

$$\mathbb{U}^{\mathsf{X}|\mathsf{D}} = \mathbb{U}\mathbb{F}^{\mathsf{X}} \tag{1.49}$$

$$= (\gamma \otimes \mathrm{Id}^D)\mathbb{T}\mathbb{F}^{\mathsf{X}} \tag{1.50}$$



$$= \tag{1.51}$$

$$= \tag{1.52}$$

Which implies $\mathsf{X} \perp\!\!\!\perp_{\mathbb{U}} \mathsf{D}$ by version (2) of conditional indpendence (Theorem 0.1.33). □

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 $\mathbb{B} : \mathsf{H} \mapsto$ Bernoulli(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 H of the coin landing on heads is as much as we can hope to know about the success of our bet. 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 1.1.19 (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 1.1.20 (Expected utility)**.** Given a utility function $u : Y \to \mathbb{R}$ and probability measures $\mu, \nu \in \Delta(\mathcal{Y})$, the *expected utility* of μ is $\mathbb{E}_\mu[u]$.

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

Definition 1.1.21 (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 1.1.22 (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

$$R(\mathsf{H}, \mathbb{U}) := -\int_X \mathbb{U}_x \mathbb{T}^{\mathsf{Y}|\mathsf{DXH}}_{\cdot,x,\mathsf{H}} u d\mathbb{T}^{\mathsf{X}|\mathsf{H}}_{\mathsf{H}}(x) \tag{1.53}$$

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 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 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 1.1.23 (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 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

1.2 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 0.1.10, but I'd rather not change his definitions if I can avoid it

put the following inside a quote environment somehow, the regular quote environment fails due to too much markup

"‘

Definition 7.1.1 (Causal Model) A causal model is a triple $M = \langle U, V, F \rangle$, where:

(i) U is a set of *background* variables, (also called *exogenous*), that are determined by factors outside the model;

(ii) V is a set $\{V_1, V_2, ..., V_n\}$ of variables, called *endogenous*, that are determined by variables in the model – that is, variables in $U \cup V$;

(iii) F is a set of functions $\{f_1, f_2, ..., f_n\}$ such that each f_i is a mapping from (the respective domains of) $U_i \cup PA_i$ to V_i, where $Ui \subseteq U$ and $PA_i \subseteq V \setminus V_i$ and the entire set F forms a mapping from U to V. In other words, each f_i in

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

assigns a value to V_i that depends on (the values of) a select set of variables in $V \cup U$, and the entire set F has a unique solution $V(u)$.

Definition 7.1.2 (Submodel) Let M be a causal model, X a set of variables in V, and x a particular realization of X. A submodel M_x of M is the causal model

$$M_x = \{U, V, F_x\},$$

where

$$F_x = \{f_i : V_i \notin X\} \cup \{X = x\}.$$

Definition 7.1.3 (Effect of Action) Let M be a causal model, X a set of variables in V, and x a particular realization of X. The effect of action $do(X = x)$ on M is given by the submodel M_x

Definition 7.1.4 (Potential Response) Let X and Y be two subsets of variables in V. The potential response of Y to action $do(X = x)$, denoted $Y_x(u)$, is the solution for Y of the set of equations F_x, that is, $Y_x(u) = Y_{M_x}(u)$.

Definition 7.1.6 (Probabilistic Causal Model) A probabilistic causal model is a pair $\langle M, P(u) \rangle$, where M is a causal model and $P(u)$ is a probability function defined over the domain of U. "'

Implicitly, Definition 7.1.3 proposes a set of "actions" that have "effects" given by M_x. It's not entirely clear what this set of actions should be – the definition seems to suggest that there is an action for each "realization" of each variable in V, which would imply that the set of actions corresponds to the range of V. For the following discussion, we will call the set of actions D, whatever it actually contains (we have deliberately chosen to use the same letter as we use to represent choices or actions in see-do models).

Given D, Definition 7.1.3 appears to define a function $h : \mathcal{M} \times D \to \mathcal{M}$, where $\mathcal{M}$ is the space of causal models with background variables U and endogenous variables V, such that for $M \in \mathcal{M}$, $do(X = x) \in D$, $h(M, do(X = x)) = M_x$.

Definition 7.1.4 then appears to define a function $Y.(\cdot) : D \times U \to Y$ (distinct from Y, which appears to be a function $U \to$ something) and calls $Y.(\cdot)$ the "potential response". We could always consider the variable $\mathsf{V} := \underline{\otimes}_{i \in [n]} \mathsf{V}_i$ and define the "total potential response" $\mathbf{g} := \mathsf{V}.(\cdot)$, which captures the potential responses of any subset of variables in V.

From this, we might surmise that in the Pearlean view, it is necessary that a "counterfactual" or "potential response" model has a probability measure P on background variables U, a set of actions D and a *deterministic* potential response function $\mathbf{g} : D \times U \to V$.

Pearl's model also features a second deterministic function $\mathbf{f} : U \to Y$, and G is derived from F via the equation modifications permitted by D. It is straightforward to show that an arbitrary function $\mathbf{f} : U \to Y$ can be constructed from Pearl's set of functions f_i, and if D may modify the set F arbitrarily, then it appears that $\mathbf{g}$ can in principle be an arbitrary function $D \times U \to Y$ (though many possible choices would be quite unusual).

Pearl's counterfactual model seems to essentially be a deterministic map $\mathbf{g} : D \times U \to V$ along with a probability measure P on U. Putting these together and marginalising over U (as we might expect we want to do with "background variables") simply yields a consequence map $D \to \Delta(\mathcal{V})$, which doesn't seem to have any special counterfactual properties.

In order to pose counterfactual questions, Pearl introduces the idea of holding U fixed:
"

Definition 7.1.5 (Counterfactual) Let X and Y be two subsets of variables in V. The counterfactual sentence "Y would be y (in situation u), had X been x" is interpreted as the equality $Y_x(u) = y$, with $Y_x(u)$ being the potential response of Y to $X = x$.' "'

Holding U fixed allows SCM counterfactual models to answer questions about what would have happened if we had taken different actions given the same background context. For example, we can compare $Y_x(u)$ with $Y_{x'}(u)$ and interpret the comparison as telling us what would have happened in the same situation u if we did x and, at the same time, what would happen if we did x'. It is the ability to consider different actions "in exactly the same situation" that makes these models "counterfactual".

One obvious question is: does $\mathbf{g}$ have to be deterministic? While SCMs are defined in terms of deterministic functions with noise arguments, it's not clear that this is a necessary feature of counterfactual models. If $\mathbf{g}$ were properly stochastic, what is the problem with considering $\mathbf{g}(x, u)$ and $\mathbf{g}(x', u)$ to represents what would happen in a fixed situation u if I did x and if I did x' respectively? In fact, a nondeterministic $\mathbf{g}$ arguably fails to capture a key intu-

ition 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 2

See-do models and the causal modelling zoo

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 (1972) after Wald's death. See-do models equipped with a *utility function* induce a slightly generalised form of statistical decision problems, and the complete class theorm is appliccable to these models.

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

Sometimes we are interested in modelling situations where we can also make some choices that also affect the eventual consequences. For example, I might hypothesise H_1: the switch on the wall controls my light, H_2: the switch on the wall does not control my light. Then, given H_1 I can choose to toggle the switch, and I will see my light turn on, or I can choose not to toggle the switch and I will not see my light turn on. Given H_2, neither choice will result in a light turned on. Choices are clearly different to hypotheses: the choice I make depends on what I want to happen, while whether or not a hypothesis is true has no regard for my ambitions.

A "statistical model with choices" is simply a map $\mathbb{T} : D \times \mathsf{H} \to \Delta(\mathcal{E})$ for some set of choices D, hypotheses H and outcome space $(E, \mathcal{E})$. We can also distinguish two types of outcomes: *observations* which are given prior to a choice being made and *consequences* which happen after a choice is made.

Observations cannot be affected by the choices made, while consequences are not subject to this restriction. That is, observations are what we might *see* before making a choice, which depends on the hypothesis alone, and if we are lucky we may be able to invert this dependence to learn something about the hypothesis from observations. On the other hand, the consequences of what we *do* depends jointly on the hypothesis and the choice we make and we judge which choices are more desirable on the basis of which consequences we expect them to produce.

What we are studying is a family of models that generalises of statistical models to include hypotheses, choices, observations and consequences. These models are referred to as *see-do models*. Hypotheses, observations, consequences and choices are not individually new ideas. *Statistical decision problems* (Wald, 1950; Savage, 1972) 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).

2.0.1 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 0.1.4 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 2.0.1 (D-causation)**.** Given a hypothesis $h \in H$ and a consequence model $\mathbb{C} : H \times D \to \Delta(\mathcal{Y})$, random variables $\mathsf{Y}_1 : Y \times D \to Y_1$, $\mathsf{Y}_2 : Y \times D \to Y_2$ and $\mathsf{D} : Y \times D \to D$ (defined the usual way), Y_1 D-causes Y_2 iff $\mathsf{Y}_2 \perp\!\!\!\perp_{\mathbb{C}} \mathsf{D}|\mathsf{Y}_1\mathsf{H}$.

2.0.2 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 2.0.2 (Limited unresponsiveness)**.** Given states S, deterministic consequence models $\mathbb{C}_s : D \to \Delta(F)$ for each $s \in A$ and a random variables $\mathsf{Y}_1 : F \to Y_1$, $\mathsf{Y}_2 : F \to Y_2$, Y_1 is unresponsive to D in states limited by Y_2 if $\mathbb{C}^{\mathsf{Y}_2|\mathsf{SD}}_{(s,d)} = \mathbb{C}^{\mathsf{Y}_2|\mathsf{SD}}_{(s,d')} \implies \mathbb{C}^{\mathsf{Y}_1|\mathsf{SD}}_{(s,d)} = \mathbb{C}^{\mathsf{Y}_1|\mathsf{SD}}_{(s,d')}$ for all $d, d' \in D$, $s \in S$. Write $\mathsf{Y}_1 \not\rightsquigarrow_{\mathsf{Y}_2} \mathsf{D}$

Lemma 2.0.3 (Limited unresponsiveness implies D-causation)**.** *For deterministic consequence models,* $\mathsf{Y}_1 \not\rightsquigarrow_{\mathsf{Y}_2} \mathsf{D}$ *implies* Y_2 *D-causes* Y_1*.*

Proof. By the assumption of determinism, for each $s \in S$ and $d \in D$ there exists $y_1(s,d)$ and $y_2(s,d)$ such that $\mathbb{C}^{\mathsf{Y}_1\mathsf{Y}_2|\mathsf{SD}}_{s,d} = \delta_{y_1(s,d)} \otimes \delta_{y_2(s,d)}$.

By the assumption of limited unresponsiveness, for all d, d' such that $y_2(s,d) = y_2(s,d')$, $y_1(s,d) = y_1(s,d')$ also. Define $f : Y_2 \times S \to Y_1$ by $(s, y_1) \mapsto y(s, [y_1(s,\cdot)]^{-1}(y_1(s,d)))$ where $[y_1(s,\cdot)]^{-1}(a)$ is an arbitrary element of $\{d | y_1(s,d) = a\}$. For all s, d, $f(y_1(s,d), s) = y_2(s,d)$. Define $\mathbb{M} : Y_2 \times S \times D \to \Delta(\mathcal{Y}_1)$ by $(y_2, s, d) \mapsto \delta_{f(y_2,s)}$. $\mathbb{M}$ is a version of $\mathbb{C}^{\mathsf{Y}_1|\mathsf{Y}_2,\mathsf{S},\mathsf{D}}$ because, for all $A \in \mathcal{Y}_2$, $B \in \mathcal{Y}_1$, $s \in S$, $d \in D$:

$$\mathbb{C}^{\mathsf{Y}_2|\mathsf{SD}}_{(s,d)} \curlyvee (\mathbb{M} \otimes \mathrm{Id}) = \int_A \mathbb{M}(y_2', d, s; B) d\delta_{y_2(s,d)}(y_2') \tag{2.1}$$

$$= \int_A \delta_{f(y_2',s)}(B) d\delta_{y_2(s,d)}(y_2') \tag{2.2}$$

$$= \delta_{f(y_2(s,d),s)}(B)\delta_{y_2(s,d)}(A) \tag{2.3}$$

$$= \delta_{y_1(s,d)}(B)\delta_{y_2(s,d)}(A) \tag{2.4}$$

$$= \delta_{y_2(s,d)} \otimes \delta_{y_1(s,d)}(A \times B) \tag{2.5}$$

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

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

define this

$$\mathbb{C}^{\mathsf{LF}|\mathsf{DS}}_{d,[i,m]} = \delta_{(d \text{ XOR } i) \text{ AND } m} \otimes \delta_{(d \text{ XOR } i) \text{ AND } m} \tag{2.6}$$

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 $\mathsf{L} \not\leftarrow_{\mathsf{F}} \mathsf{D}$. However, suppose we take a mixture of consequence models:

$$\mathbb{C}_\gamma = \frac{1}{4}\mathbb{C}_{\cdot,[0,0]} + \frac{1}{4}\mathbb{C}_{\cdot,[0,1]} + \frac{1}{2}\mathbb{C}_{\cdot,[1,1]} \tag{2.7}$$

$$\mathbb{C}_\gamma^{\mathsf{FL}|\mathsf{D}} = \frac{1}{4}\begin{bmatrix}1&0\\0&1\end{bmatrix} \otimes \begin{bmatrix}1&0\\1&0\end{bmatrix} + \frac{1}{4}\begin{bmatrix}1&0\\0&1\end{bmatrix} \otimes \begin{bmatrix}1&0\\0&1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}0&1\\1&0\end{bmatrix} \otimes \begin{bmatrix}0&1\\1&0\end{bmatrix} \tag{2.8}$$

Then

$$[1,0]\mathbb{C}_\gamma^{\mathsf{FL}|\mathsf{D}} = \frac{1}{4}[0,1] \otimes [1,0] + \frac{1}{4}[0,1] \otimes [0,1] + \frac{1}{2}[1,0] \otimes [1,0] \tag{2.9}$$

$$[1,0]\curlyvee(\mathbb{C}_\gamma^{\mathsf{F}|\mathsf{D}} \otimes \mathbb{C}_\gamma^{\mathsf{L}|\mathsf{D}}) = (\frac{1}{2}[0,1] + \frac{1}{2}[1,0]) \otimes (\frac{1}{4}[0,1] + \frac{3}{4}[1,0]) \tag{2.10}$$

$$\implies [1,0]\mathbb{C}_\gamma^{\mathsf{FL}|\mathsf{D}} \neq [1,0]\curlyvee(\mathbb{C}_\gamma^{\mathsf{F}|\mathsf{D}} \otimes \mathbb{C}_\gamma^{\mathsf{L}|\mathsf{D}}) \tag{2.11}$$

define this

Thus under the prior γ, F does not D-cause L even though F D-causes L in all states S. The definition of D-causation was motivated by the idea that we could reduce a difficult decision problem with a large set D to a simpler problem with a smaller "effective" set of decisions by exploiting conditional independence. Even if X D-causes Y in every $\mathsf{H} \in S$, X does not necessarily D-cause Y in mixtures of states in S. For this reason, we do not say that X D-causes Y in S if X D-causes Y in every $\mathsf{H} \in S$, and in this way we differ substantially from Heckerman and Shachter (1995).

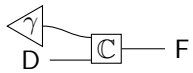
Instead, we simply extend the definition of D-causation to mixtures of hypotheses: if $\gamma \in \Delta(\mathsf{H})$ is a mixture of hypotheses, define $\mathbb{C}_\gamma := (\gamma \otimes \mathbf{Id})\mathbb{C}$. Then X D-causes Y relative to γ iff $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{C}_\gamma}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

Theorem 2.0.4 shows that under some conditions, D-causation can hold for arbitrary mixtures over subsets of the hypothesis class H.

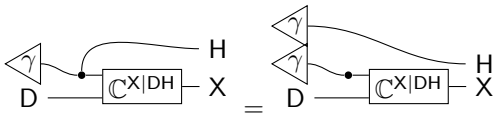
Theorem 2.0.4 (Universal D-causation)**.** *If* $\mathsf{X} \perp\!\!\!\perp \mathsf{H}|\mathsf{D}$ *for all* $\mathsf{H}, \mathsf{H}' \in S \subset \mathsf{H}$ *and* X *D-causes* Y *in all* $\mathsf{H} \in S$*, then* X *D-causes* Y *with respect to all mixed consequence models* $\mathbb{C}_\gamma$ *for all* $\gamma \in \Delta(\mathsf{H})$ *with* $\gamma(S) = 1$*.*

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

$$\mathbb{C}_\gamma := \tag{2.12}$$

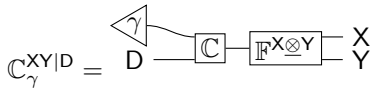


Because $\mathbb{C}_{\mathsf{H}}^{\mathsf{X}|\mathsf{D}} = \mathbb{C}_{\mathsf{H}'}^{\mathsf{X}|\mathsf{D}}$ for all $\mathsf{H}, \mathsf{H}' \in \mathsf{H}$, we have

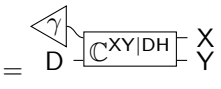


(2.13)

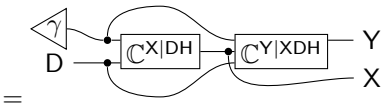
Also



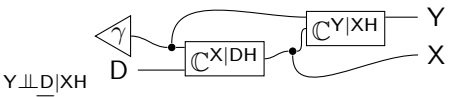
(2.14)



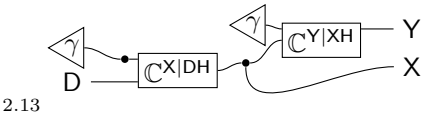
(2.15)



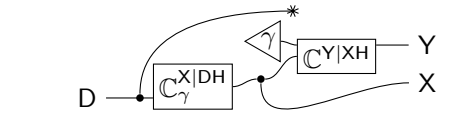
(2.16)



(2.17)



(2.18)



(2.19)

Equation 2.19 establishes that $(\gamma \otimes \mathbf{Id}_X \otimes {}^*\!\uparrow_D)\mathbb{C}^{\mathsf{Y|XH}}$ is a version of $\mathbb{C}^{\mathsf{Y|XD}}_\gamma$, and thus $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{C}_\gamma} \mathsf{D|X}$.

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

$$\mathsf{H} \perp\!\!\!\perp \mathsf{D} \wedge \mathsf{H} \perp\!\!\!\perp \mathsf{X|D} \implies \mathsf{H} \perp\!\!\!\perp \mathsf{XD} \tag{2.20}$$
$$\implies \mathsf{H} \perp\!\!\!\perp \mathsf{D|X} \tag{2.21}$$
$$\mathsf{D} \perp\!\!\!\perp \mathsf{H|X} \wedge \mathsf{D} \perp\!\!\!\perp \mathsf{Y|XH} \implies \mathsf{D} \perp\!\!\!\perp \mathsf{Y|X} \tag{2.22}$$
$$\implies \mathsf{Y} \perp\!\!\!\perp \mathsf{D|X} \tag{2.23}$$

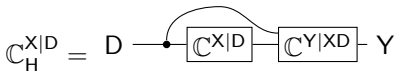
□

2.0.3 Properties of D-causation

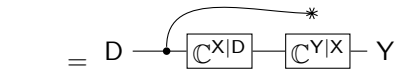
If X D-causes Y relative to $\mathbb{C}_\mathsf{H}$, then the following holds:

(2.24)

This follows from version (2) of Definition 0.1.35:



(2.25)



(2.26)

(2.27)

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

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

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

2.0.4 Decision sequences and parallel decisions

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

Chapter 3

Chapter 5: Inferring causes from data

References

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`.

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.

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/.

Chayakrit Krittanawong, Bharat Narasimhan, Zhen Wang, Joshua Hahn, Hafeez Ul Hassan Virk, Ann M. Farrell, HongJu Zhang, and WH Wilson Tang. Association between chocolate consumption and risk of coronary artery disease: a systematic review and meta-analysis:. *European Journal of Preventive Cardiology*, July 2020. doi: 10.1177/2047487320936787. URL http://journals.sagepub.com/doi/10.1177/2047487320936787. Publisher: SAGE PublicationsSage UK: London, England.

Finnian Lattimore and David Rohde. Replacing the do-calculus with Bayes rule. *arXiv:1906.07125 [cs, stat]*, December 2019. URL http://arxiv.org/abs/1906.07125. arXiv: 1906.07125.

L. Le Cam. Comparison of Experiments - A Short Review.pdf. *IMS Lecture Notes - Monograph Series*, 30, 1996.

Dennis Nilsson and Steffen L. Lauritzen. Evaluating Influence Diagrams using LIMIDs. *arXiv:1301.3881 [cs]*, January 2013. URL http://arxiv.org/abs/1301.3881. arXiv: 1301.3881.

Naomi Oreskes and Erik M. Conway. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Climate Change: How a Handful of Scientists ... Issues from Tobacco Smoke to Global Warming*. Bloomsbury Press, New York, NY, June 2011. ISBN 978-1-60819-394-3.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 1 edition edition, May 2018. ISBN 978-0-465-09760-9.

Robert N. Proctor. The history of the discovery of the cigarettelung cancer link: evidentiary traditions, corporate denial, global toll. *Tobacco Control*, 21(2):87–91, March 2012. ISSN 0964-4563, 1468-3318. doi: 10.1136/tobaccocontrol-2011-050338. URL `https://tobaccocontrol.bmj.com/content/21/2/87`. Publisher: BMJ Publishing Group Ltd Section: The shameful past.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

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. *Foundations of Statistics*. Dover Publications, New York, revised edition edition, June 1972. ISBN 978-0-486-62349-8.

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.

Abraham Wald. *Statistical decision functions*. Statistical decision functions. Wiley, Oxford, England, 1950.

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: