Causal Statistical Decision Theory|What are interventions?

David Johnston

May 18, 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 $P^{\mathsf{X}|\mathsf{ZY}}$

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.

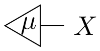
Indiscrete spaces play a key role in string diagrams. An indiscrete space is any one element measurable space $(\{*\}, \{\emptyset, \{*\}\})$ which admits the unique probability measure $\mu : \{\emptyset, \{*\}\} \to (0, 1)$ given by $\mu(\emptyset) = 0$, $\mu(\{*\}) = 1$. Any probability measure $\mu \in \Delta(\mathcal{X})$ 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:

$$X -\boxed{\mathbb{M}}- Y \tag{6}$$

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

Probability measures are kernels with an indiscrete domain $\mu \in \Delta(\mathcal{X})$ can be written as triangles:



(7)

Note that Eq 7 technically represents a Markov kernel $\mu' : \{*\} \to \Delta(\mathcal{X})$, but for our purposes this distinction isn't practically important.

We do *not* define kernel-function products for string diagrams. While kernel-kernel products always yield Markov kernels as a result, and measure-kernel products can be reinterpreted as kernel-kernel products, kernel-function products do not admit such a reinterpretation. Cho and Jacobs (2019) defines the operation of *conditioning* using kernel-function products, but this will take extra work to incorporate into our notation which hasn't yet been done.

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

$$\mathbb{M}\mathbb{N} := X \text{ —}\boxed{\mathbb{M}}\text{—}\boxed{\mathbb{N}}\text{— } Z \tag{8}$$

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

$$\mu\mathbb{M} := \lhd\!\mu\text{—}\boxed{\mathbb{M}}\text{— } Z \tag{9}$$

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

$$\mathbb{M} \otimes \mathbb{O} := \begin{array}{l} X \text{ —}\boxed{\mathbb{M}}\text{— } Y \\ Z \text{ —}\boxed{\mathbb{O}}\text{— } W \end{array} \tag{10}$$

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:

$$\mathrm{Id}^X := X \text{ ——————— } X \tag{11}$$

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:

$$X \times Y \cong \mathrm{Id}^X \otimes \mathrm{Id}^Y := \begin{array}{l} X \text{ — } X \\ Y \text{ — } Y \end{array} \tag{12}$$

$$= X \times Y \text{ —— } X \times Y \tag{13}$$

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:

$$X \text{ —}\boxed{\mathbb{L}}\begin{array}{l}\text{— } Y \\ \text{— } Z\end{array} \tag{14}$$

$$\equiv \tag{15}$$

$$X \text{ —}\boxed{\mathbb{L}}\text{—— } Y \times Z \tag{16}$$

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:

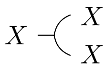
$$\mathbf{Id} : x \mapsto \delta_x \equiv X \text{ — } X \tag{17}$$

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:

$$\curlyvee : x \mapsto \delta_{(x,x)} \equiv \tag{25}$$

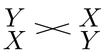


The copy map "copies" its arguments to kernels or under the right product:

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

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

$$\sigma := (x,y) \to \delta_{(y,x)} \equiv \tag{28}$$



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

$$
\begin{aligned}
(\sigma\mathbb{K})_{y,x}(A) &= \int_{(} X \times Y)\mathbb{K}_{x',y'}(A)d\sigma_{(y,x)}(x',y') \quad = \int_{(} X \times Y)\mathbb{K}_{x',y'}(A)d\delta_{(x,y)}(x',y') \tag{29}\\
&= \mathbb{K}_{x,y}(A) \tag{30}\\
(\mathbb{L}\sigma)_z(B \times A) &= \int_{X\times Y} \sigma_{x',y'}(B\times A)d\mathbb{L}_z(x',y') \tag{31}\\
&= \int_{X\times Y} \delta_{(y',x')}(B\times A)d\mathbb{L}_z(x',y') \tag{32}\\
&= \mathbb{L}_z(A\times B) \tag{33}
\end{aligned}
$$

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

$$\ast : x \mapsto \delta_* \equiv X \longrightarrow\!\ast \tag{34}$$

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 $a \in V$, $B \in \mathcal{X}$:

$$
\begin{aligned}
(\mathbb{M}\mathbb{F}^g)_a(B) &= \int_W \delta_{g(y)}(B)d\mathbb{M}_a(y) \tag{35}\\
&= \int_W \delta_y(g^{-1}(B))d\mathbb{M}_a(y) \tag{36}\\
&= \int_{g^{-1}(B)} d\mathbb{M}_a(y) \tag{37}\\
&= (g_{\#}\mathbb{M})_a(B) \tag{38}
\end{aligned}
$$

□

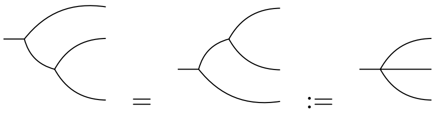
Working With String Diagrams

todo:

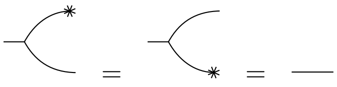
- Infinite copy map
- De Finetti's representation theorem

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.

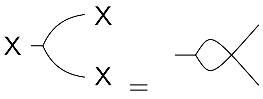
Axioms of Symmetric Monoidal Categories For the following, we either omit labels or label diagrams with their domain and codomain spaces, as we are discussing identities of kernels rather than identities of components of a condtional probability space. Recalling the unique Markov kernels defined above,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



(39)

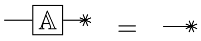


(40)



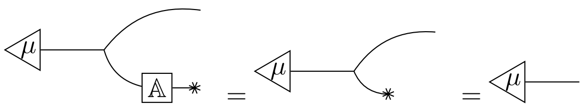
(41)

The discard map ∗ can "fall through" any Markov kernel:



(42)

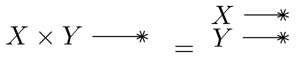
Combining 40 and 42 we can derive the following: integrating $\mathbb{A} : X \to \Delta(\mathcal{Y})$ with respect to $\mu \in \Delta(\mathcal{X})$ and then discarding the output of $\mathbb{A}$ leaves us with μ:



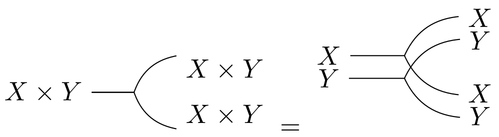
(43)

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 $\curlyvee$:

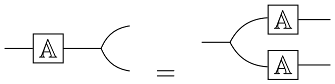


(44)



(45)

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



(46)

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 46 holds iff $\mathbb{A}$ is deterministic.*

Examples

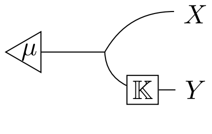
Given $\mu \in \Delta(X), \mathbb{K} : X \to \Delta(Y)$, $A \in \mathcal{X}$ and $B \in \mathcal{Y}$:

$$A \times B \mapsto \int_A \mathbb{K}(x;B)d\mu(x) \tag{47}$$

$$\equiv \tag{48}$$

$$\mu\curlyvee(\mathbf{Id}_X \otimes \mathbb{K}) \tag{49}$$

$$\equiv \tag{50}$$

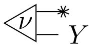


(51)

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

$$\nu(\ast \otimes \mathrm{Id}^Y) = \tag{52}$$

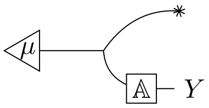


$$\begin{aligned}
\nu(\ast \otimes \mathrm{Id}^Y)(B) &:= \nu(\ast \otimes \mathrm{Id}^Y)(B \times \{*\}) &(53)\\
&= \int_{X\times Y} \mathrm{Id}^Y_y(B)\ast_x(\{*\})d\nu(x,y) &(54)\\
&= \int_{X\times Y} \delta_y(B)\delta_*(\{*\})d\nu(x,y) &(55)\\
&= \int_{X\times B} d\nu(x,y) &(56)\\
&= \nu(X\times B) &(57)\\
&= \nu^{\mathsf{Y}}(B) &(58)
\end{aligned}$$

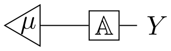
Thus the action of the erasing wire "X" is equivalent to marginalising over "X".

Consider the result of marginalising 51 over "X":

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



$$= \tag{60}$$



0.1.4 Random Variables

The summary of this section is:

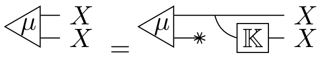
- Random variables are usually defined as measurable functions on a *probability space*
- It's possible to define them as measurable functions on a *Markov kernel space* instead
- It is useful to label wires with random variable names instead of names of spaces

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.

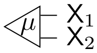
Suppose that given a particular distribution over decision variables, a probability distribution over the decision variables and random variables is obtained. Such a model is described by a Markov kernel rather than a probability distribution. We therefore investigate *Markov kernel spaces*.

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:



(61)

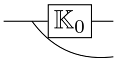
Suppose that we also assign the names X_1 to the upper output wire and X_2 to the lower output wire in the diagram of μ:



(62)

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 $(D, \mathcal{D})$. We suppose that $\mathbb{K}$ is such that there exists a *fundamental kernel* $\mathbb{K}_0$ satisfying

$\mathbb{K} :=$ 

(63)

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 $\Omega \times \{*\} \cong \Omega$.

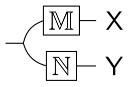
Definition 0.1.8 (Random variable)**.** Given a Markov kernel space $(\mathbb{K}, D, \Omega)$, a random variable X is a measurable function $\Omega \times D \to E$ for arbitrary measurable E.

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 $\mathsf{D} : (x, d) \mapsto d$.

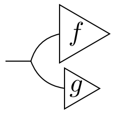
Unlike random variables on probability spaces, random variables on Markov kernel spaces do not generally have unique marginal distributions. An analogous operation of *marginalisation* can be defined, but the result is generally a Markov kernel. We will define marginalisation via coupled tensor products.

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:

$$\mathbb{M}\underline{\otimes}\mathbb{N} := \tag{64}$$



$$f\underline{\otimes}g := \tag{65}$$



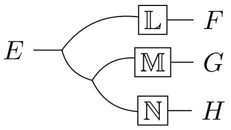
The operation denoted by $\underline{\otimes}$ is associative (Lemma 0.1.11), so we can without ambiguity write $f\underline{\otimes}g\underline{\otimes}h = (f\underline{\otimes}g)\underline{\otimes}h = f\underline{\otimes}(g\underline{\otimes}h)$ for finite groups of functions or Markov kernels sharing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

The notation $\underline{\otimes}_{i\in[N]} f_i$ is taken to mean $f_1\underline{\otimes}f_2\underline{\otimes}...\underline{\otimes}f_N$.

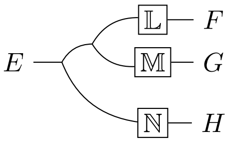
Lemma 0.1.11 ($\underline{\otimes}$ is associative)**.** *For Markov kernels* $\mathbb{L} : E \to \delta(\mathcal{F})$*,* $\mathbb{M} : E \to \delta(\mathcal{G})$ *and* $\mathbb{N} : E \to \delta(\mathcal{H})$*,* $(\mathbb{L}\underline{\otimes}\mathbb{M})\underline{\otimes}\mathbb{N} = \mathbb{L}\underline{\otimes}(\mathbb{M}\underline{\otimes}\mathbb{N})$*.*

Proof.

$$\mathbb{L}\underline{\otimes}(\mathbb{M}\underline{\otimes}\mathbb{N}) = \tag{66}$$



$$= \tag{67}$$



$$= (\mathbb{L}\underline{\otimes}\mathbb{M})\underline{\otimes}\mathbb{N} \tag{68}$$

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 $\mathbb{K}^{\mathsf{X}|\mathsf{D}} := \mathbb{K}\mathbb{F}^{\mathsf{X}}$.

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 $\mathbb{K}^{\mathsf{XY|D}}$ 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.*

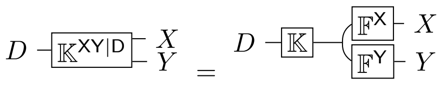
Given $\mathsf{X} : \Omega \to G$ *and* $\mathsf{Y} : \Omega \to H$*,* $\mathbb{F}^{\mathsf{X}\underline{\otimes}\mathsf{Y}} = \mathbb{F}^{\mathsf{X}}\underline{\otimes}\mathbb{F}^{\mathsf{Y}}$

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

$$\begin{aligned} \mathbb{F}^{\mathsf{X}\underline{\otimes}\mathsf{Y}}(a; B \times C) &= \delta_{\mathsf{X}(a),\mathsf{Y}(a)}(B \times C) && (69)\\ &= \delta_{\mathsf{X}(a)}(B)\delta_{\mathsf{Y}(a)}(C) && (70)\\ &= (\delta_{\mathsf{X}(a)} \otimes \delta_{\mathsf{Y}(a)})(B \times C) && (71)\\ &= \mathbb{F}^{\mathsf{X}}\underline{\otimes}\mathbb{F}^{\mathsf{Y}} && (72) \end{aligned}$$

Equality follows from the monotone class theorem. □

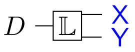
Corollary 0.1.15. *Given a Markov kernel space* $(\mathbb{K}, \Omega, D)$ *and random variables* $\mathsf{X} : \Omega \times D \to X$*,* $\mathsf{Y} : \Omega \times D \to Y$*, the following holds:*


D
$\mathbb{K}^{\mathsf{XY|D}}$
X
Y
=
D
$\mathbb{K}$
$\mathbb{F}^{\mathsf{X}}$
X
$\mathbb{F}^{\mathsf{Y}}$
Y

(73)

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 $\mathbf{L}$:

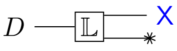


(74)

is *valid* iff

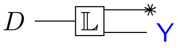
$$\mathbb{L} = \mathbb{K}_{\mathsf{XY|D}} \tag{75}$$

and



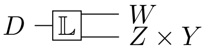
$$= \mathbb{K}^{\mathsf{X}|\mathsf{D}} \tag{76}$$

and



$$= \mathbb{K}^{\mathsf{Y}|\mathsf{D}} \tag{77}$$

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

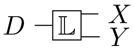


(78)

For *this* diagram, properties 76 and 77 do not hold, even though 75 does.

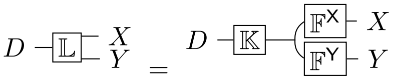
Lemma 0.1.17 (Output label assignments exist)**.** *Given Markov kernel space* $(\mathbb{K}, D, \Omega)$*, random variables* $\mathsf{X} : \Omega \times D \to X$ *and* $\mathsf{Y} : \Omega \times D \to Y$ *then there exists a diagram of* $\mathbb{L} := \mathbb{K}^{\mathsf{XY}|\mathsf{D}}$ *with a valid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



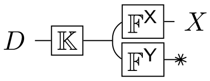
(79)

By Corollary 0.1.15, we have

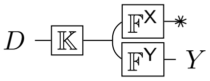


(80)

Therefore



$$
\begin{aligned}
&= \mathbb{K}\mathbb{F}^{\mathsf{X}} && (81)\\
&= \mathbb{K}^{\mathsf{X}|\mathsf{D}} && (82)
\end{aligned}
$$



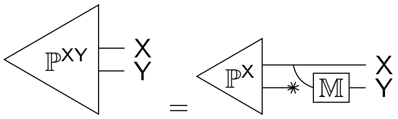
$$= \mathbb{K}\mathbb{F}^{\mathsf{Y}} \tag{83}$$

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

□

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

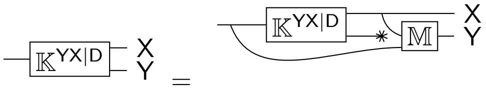
Definition 0.1.18 (Disintegration)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$, and random variables X and Y, we say that $\mathbb{M} : E \to \Delta(\mathcal{F})$ is a Y *on* X *disintegration* of $\mathbb{P}$ iff



(85)

$\mathbb{M}$ 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 85 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 *on* DX *disintegration* of $\mathbb{K}^{\mathsf{YX}|\mathsf{D}}$ iff



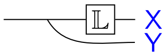
(86)

Write $\mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$ for the set of kernels satisfying 86 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.

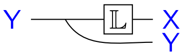
If an input wire is connected to an output wire and that output wire has a valid label X, then it is valid to label the input wire with X.

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



(87)

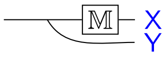
i.e. if $\mathbb{L} \in \mathbb{P}^{\{\mathsf{XY}|\mathsf{Y}\}}$, then the following is a valid input label:



(88)

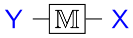
An input wire in a diagram for $\mathbb{M}$ may be labeled X *if and only if* copy and identity maps can be inserted to yield a diagram in which the input wire labeled X is connected to an output wire with valid label X.

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



$$\in \mathbb{P}^{\{\mathsf{XY}|\mathsf{Y}\}} \qquad (89)$$

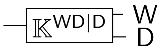
and hence the output labels are valid. Diagram 89 is constructed by taking the product of the copy map with $\mathbb{M} \otimes \mathbf{Id}$. Thus it is valid to label $\mathbb{M}$ with



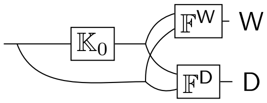
(90)

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 $\mathbb{K}$:

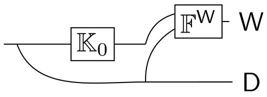


$=$



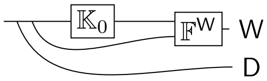
(91)

$=$



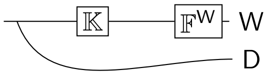
(92)

$=$



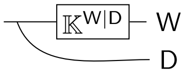
(93)

$=$



(94)

$=$

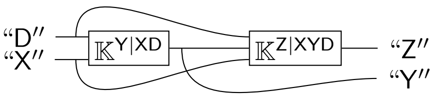


(95)

□

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*,*


"D″
"X″
$\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}$
$\mathbb{K}^{\mathsf{Z}|\mathsf{XYD}}$
"Z″
"Y″

$$\in \mathbb{K}^{\{\mathsf{ZY}|\mathsf{XD}\}} \qquad (96)$$

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

$$(d, x; A, B) \mapsto \int_A \mathbb{K}^{\mathsf{Z}|\mathsf{XYD}}_{(x,y,d)}(B) d\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}_{(x,d)}(y) \in \mathbb{K}^{\{\mathsf{ZY}|\mathsf{XD}\}} \qquad (97)$$

Proof.

write this up

□

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

Theorem 0.1.22 (Disintegration existence - probability space)**.** *Given a probability measure* $\mu \in \Delta(\mathcal{X} \otimes \mathcal{Y})$*, if* $(F, \mathcal{F})$ *is standard then a disintegration* $\mathbb{K} : X \to \Delta(\mathcal{Y})$ *exists (Çinlar, 2011).*

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

For Markov kernel spaces, we make the simplifying assumption that the domain space D is a discrete space. Given this assumption, there exists a positive definite probability $\mu \in \Delta(\mathcal{D})$. That is, for every $d \in D$, $\mu(\{d\}) > 0$. Given this assumption, for every Markov kernel space $(\mathbb{K}, D, \Omega)$ there is a probability space $(\mathbb{P}, \Omega \times D)$ such that $\mathbb{K}$ can be uniquely defined as a disintegration of $\mathbb{P}$. For uncountable D, even if it is standard measurable, this is not possible (Hájek, 2003).

Definition 0.1.23 (Relative probability space)**.**

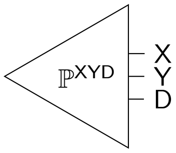
better name

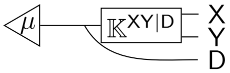
Given a Markov kernel space $(\mathbb{K}, D, \Omega)$ and a positive definite measure $\mu \in \Delta(\mathcal{D})$, $(\mu\mathbb{K}, \Omega \times D)$ is a *relative* probability space.

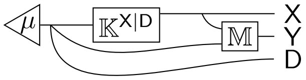
For any random variable $\mathsf{X} : \Omega \times D \to X$ on $(\mathbb{K}, D, \Omega)$, its relative on $(\mu\mathbb{K}, \Omega \times D)$ is given by the same measurable function, and we give it the same name X.

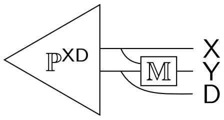
Lemma 0.1.24 (Agreement of disintegrations)**.** *Given a Markov kernel space* $(\mathbb{K}, D, \Omega)$*, any relative probability space* $(\mu\mathbb{K}, \Omega \times D)$ *and any random variables* $\mathsf{X} : \Omega \times D \to X$*,* $\mathsf{Y} : \Omega \times D \to Y$*,* $\mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}} = (\mu\mathbb{K})^{\{\mathsf{Y}|\mathsf{XD}\}}$ *(note that this set equality).*

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



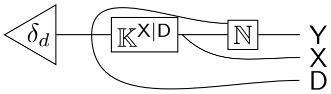
$=$  (98)

$=$  (99)

$=$  (100)

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

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

$\mathbb{Q} :=$  (101)

$= \mathbb{K}_d^{\mathsf{XYD|D}}$ (102)

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

$$\mathbb{Q}(A \times B \times \{d\}) = \int_{X^2} \int_X \int_{D^3} \mathbb{N}_{d'',x'}(A)\mathbf{Id}_{x''}(B)\mathbf{Id}_{d'''}(\{d\})d\curlyvee_d(d',d'',d''')d\mathbb{K}_{d'}^{\mathsf{X|D}}(x)d\curlyvee_x(x',x'') \tag{103}$$

$$= \delta_d(\{d\}) \int_X \mathbb{N}_{d,x}(A)\delta_x(B)d\mathbb{K}_d^{\mathsf{X|D}}(x) \tag{104}$$

$$= \frac{1}{\mu(\{d\})} \int_{\{d\}} d\mu(d') \int_X \mathbb{N}_{d,x}(A)\delta_x(B)d\mathbb{K}_d^{\mathsf{X|D}}(x) \tag{105}$$

$$= \frac{1}{\mu(\{d\})} \int_D \int_X \mathbb{N}_{d,x}(A)\delta_{d'}(\{d\})\delta_x(B)d\mathbb{K}_d^{\mathsf{X|D}}(a)d\mu(d') \tag{106}$$

$$= \frac{1}{\mu(\{d\})} \int_D \int_X \mathbb{N}_{d,x}(A)\delta_{d'}(\{d\})\delta_x(B)d\mathbb{K}_{d'}^{\mathsf{X|D}}(a)d\mu(d') \tag{107}$$

$$= \frac{1}{\mu(\{d\})} \mathbb{P}^{\mathsf{XYD}}(A \times B \times \{d\}) \tag{108}$$

$$= \frac{1}{\mu(\{d\})} \int_D \mathbb{K}_{d'}^{\mathsf{XYD|D}}(A \times B \times \{d\})d\mu(d') \tag{109}$$

$$= \frac{1}{\mu(\{d\})} \int_D \mathbb{K}_{d'}\mathsf{XY|D}(A \times B)\delta_{d'}(\{d\})d\mu(d') \tag{110}$$

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

$$= \mathbb{K}_d^{\mathsf{XY|D}}(A \times B)\delta_d(\{d\}) \tag{112}$$

$$= \int_D \mathbb{K}_{d'}^{\mathsf{XY}}(A \times B)\delta_{d''}(\{d\})d\curlyvee_d(d',d'') \tag{113}$$

$$= \mathbb{K}_d^{\mathsf{XYD|D}}(A \times B \times \{d\}) \tag{114}$$

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

Thus any kernel conditional probability $\mathbb{K}^{\mathsf{Y|XD}}$ can equally well be considered a regular conditional probability $\mathbb{P}^{\mathsf{Y|XD}}$ 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 $\mathbb{P}^{\mathsf{Y|X}}$ that is *not* conditioned on D is undefined in the kernel space $(\mathbb{K}, D, \Omega)$.

Conditional Independence

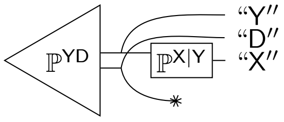
Definition 0.1.25 (Kernels constant in an argument)**.** Given a kernel $(\mathbb{K}, D, \Omega)$ and random variables Y and X, we say a verstion of the disintegration $\mathbb{K}^{\mathsf{Y|XD}}$ is constant in D if for all $x \in X$, $d, d' \in D$, $\mathbb{K}^{\mathsf{Y|XD}}_{(x,d)} = \mathbb{K}^{\mathsf{Y|XD}}_{(x,d')}$.

Definition 0.1.26 (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:

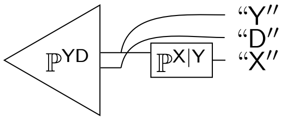
Almost sure equality

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 0.1.27 (Definitions are equivalent)**.** *(1) $\Longrightarrow$ (2): By Lemma 0.1.24,* $\mathbb{P}^{\{\mathsf{Y}|\mathsf{XD}\}} = \mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}}$*. Thus it is sufficient to show that* $\mathbb{P}^{\mathsf{X}|\mathsf{Y}} \otimes *$ *is a version of* $\mathbb{P}^{\{\mathsf{X}|\mathsf{YD}\}}$*.*

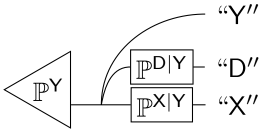

$\mathbb{P}^{\mathsf{YD}}$
$\mathbb{P}^{\mathsf{X}|\mathsf{Y}}$
“Y″
“D″
“X″
$*$

$=$


$\mathbb{P}^{\mathsf{YD}}$
$\mathbb{P}^{\mathsf{X}|\mathsf{Y}}$
“Y″
“D″
“X″

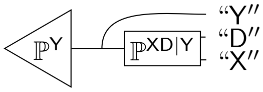
(115)

$=$


$\mathbb{P}^{\mathsf{Y}}$
$\mathbb{P}^{\mathsf{D}|\mathsf{Y}}$
$\mathbb{P}^{\mathsf{X}|\mathsf{Y}}$
“Y″
“D″
“X″

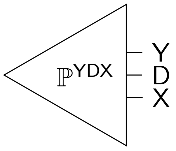
(116)

$\overset{condition(1)}{=}$


$\mathbb{P}^{\mathsf{Y}}$
$\mathbb{P}^{\mathsf{XD}|\mathsf{Y}}$
“Y″
“D″
“X″

(117)

$=$



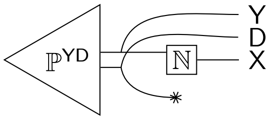
(118)

(2) $\Longrightarrow$ (3)

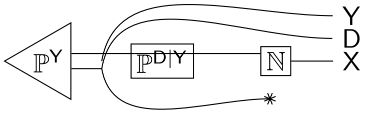
$\mathbb{P}^{\mathsf{X}|\mathsf{Y}} \otimes *_D$ *is a version of* $\mathbb{K}^{\{\mathsf{X}|\mathsf{YD}\}}$ *by assumption, and is clearly constant in* D*.*

(3) $\Longrightarrow$ (1)

By lemma 0.1.24, 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*



$=$


$\mathbb{P}^{\mathsf{Y}}$
$\mathbb{P}^{\mathsf{D}|\mathsf{Y}}$
$\mathbb{N}$
Y
D
X
$*$

(119)

Definition 0.1.28 (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 0.1.29 (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

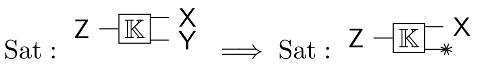
- $\mathbb{P}^{\mathsf{XZ}|\mathsf{Y}} \sim \mathbb{P}^{\mathsf{X}|\mathsf{Y}} \underline{\otimes} \mathbb{P}^{\mathsf{Z}|\mathsf{Y}}$
- For any version of $\mathbb{P}^{\{\mathsf{X}|\mathsf{Y}\}}$, $\mathbb{P}^{\mathsf{X}|\mathsf{Y}} \otimes *_Z$ is a version of $\mathbb{K}^{\{\mathsf{X}|\mathsf{YZ}\}}$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

Lemma 0.1.30 (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.*

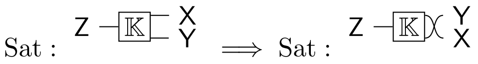
Fix some Markov kernel space $(\mathbb{K}, D, \Omega)$ *and random variables* X, Y, Z *taking values in* X, Y, Z *respectively.* Sat : *indicates that a labeled diagram satisfies definitions 0.1.16 and 0.1.19 with respect to* $(\mathcal{K}, D, \Omega)$ *and* X, Y, Z*. The following always holds:*

$$\text{Sat} : \mathsf{X} - \mathsf{X} \tag{120}$$

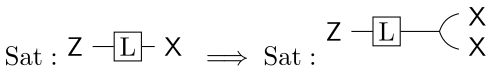
and the following implications hold:



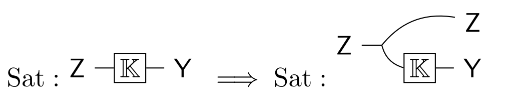
(121)



(122)



(123)



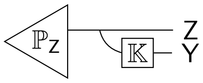
(124)

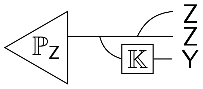
Proof.

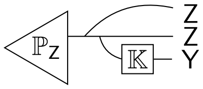
- Id_X is a version of $\mathbb{P}_{\mathsf{X}|\mathsf{X}}$ for all $\mathbb{P}$; $\mathbb{P}_{\mathsf{X}}\mathrm{Id}_X = \mathbb{P}_{\mathsf{X}}$
- $\mathbb{K}\mathrm{Id} \otimes *)(w; A) = \int_{X \times Y} \delta_x(A) \mathbb{1}_Y(y) d\mathbb{K}_w(x, y) = \mathbb{K}_w(A \times Y) = \mathbb{P}_{\mathsf{X}|\mathsf{Z}}(w;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

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

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

$\mathbb{P}_{\mathsf{ZY}} =$  (125)

$\mathbb{P}_{\mathsf{ZZY}} =$  (126)

$=$  (127)

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:

$$\text{Sat} : \mathsf{Z} -\!\boxed{\mathbb{K}}\!- \mathsf{Y}\;,\; \mathsf{Y} -\!\boxed{\mathbb{L}}\!- \mathsf{X} \implies \text{Sat} : \mathsf{Z} -\!\boxed{\mathbb{K}}\!-\!\boxed{\mathbb{L}}\!- \mathsf{X} \tag{128}$$

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

Chapter 3: 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

The basic claim of this chapter is that see-do models are the basic type of thing that everyone who is studying “causal inference” is working with, even if they don’t know it themselves

Consider the following problem: you are presented with a collection H of hypotheses about how the world might function and a vector x of observational data which you know could have taken values in some space X. You want to determine which hypothesis $\mathsf{H} \in \mathsf{H}$ best describes the world. However you ultimately solve the problem, the next step you take will probably be to determine for each $\mathsf{H} \in \mathsf{H}$ a probability distribution $\mathbb{P}_{\mathsf{H}} \in \Delta(\mathcal{X})$ that indicates how likely you would be to observe the various elements of X were H in fact the case. This is a *statistical model* – an indexed set of probability distributions $\{\mathbb{P}_{\mathsf{H}} | \mathsf{H} \in \mathsf{H}\}$. Statistical models are ubiquitous in the field of statistics – they are found in statistical decision theory where the elements of H are typically called “states”(Wald, 1950), in Bayesian inference where the elements of H may be called “parameters” (Freedman, 1963) and in frequentist inference where elements of H they may be called “hypotheses” (Fisher, 1992).

These different approaches to statistics may have different notions of what the “best hypothesis” H is, may employ different estimation methods and may not even agree about what “distributed according to $\mathbb{P}_{\mathsf{H}}$” means. Nonetheless, the interpretation of the statistical model in each case is roughly the same: supposing $\mathsf{H} \in \mathsf{H}$ is true, the data will be distributed according to $\mathbb{P}_{\mathsf{H}}$. A statistical model takes a hypothesis and tells you what you are likely to *see*.

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).

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 counterfacutla 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).

1.1 Definition

Terminology question: The variables H and D aren't necessarily *random* in the commonsense understanding of the word. They're also defined on a *kernel space* rather than a *probability space.* They're currently called random variables simply by virtue of being measurable functions on the outcome space. I'm not a huge fan of "quasirandom variables", but it does capture the idea that these things are very similar to random variables but not exactly the same.

Definition 1.1.1 (See-Do model)**.** A *see-do model* $\langle \mathbb{T}, \mathsf{H}, \mathsf{D}, \mathsf{X}, \mathsf{Y} \rangle$ is a kernel space (Definition 0.1.7) $(\mathbb{T}, H \times D, X \times Y)$ along with four random variables: the *hypothesis* $\mathsf{H} : H \times D \times X \times Y \to H$, the *choice* $\mathsf{D} : H \times D \times X \times Y \to D$, the *observations* $\mathsf{X} : H \times D \times X \times Y \to X$ and the *consequences* $\mathsf{Y} : H \times D \times X \times Y \to Y$, all given by the obvious projection maps.

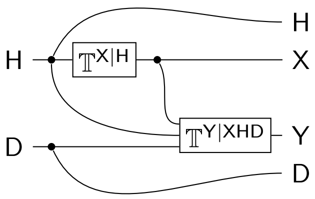
The spaces H, D, X and Y are the hypothesis, choice, observation and consequence spaces respectively.

A see-do model has the additional property that, holding the hypothesis fixed, the observations are independent of the choices - i.e. $\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$. We require that $H \times D$ is countable.

Theorem 1.1.2 (Observation and Consequence models)**.** *Any see-do model* $(\mathbb{T}, \mathsf{H}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ *can be uniquely represented by the following pair of Markov kernels:*

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

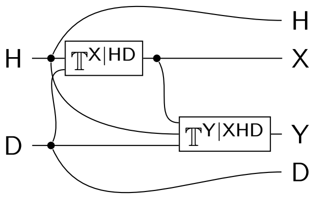
Furthermore


H
H
$\mathbb{T}^{\mathsf{X}|\mathsf{H}}$
X
$\mathbb{T}^{\mathsf{Y}|\mathsf{XHD}}$
Y
D
D

$$\mathbb{T} = \tag{1.1}$$

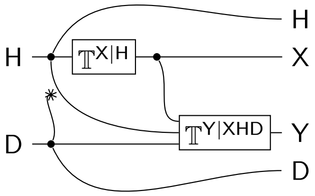
Maybe moves proofs out of main text

Proof. By 0.1.1,

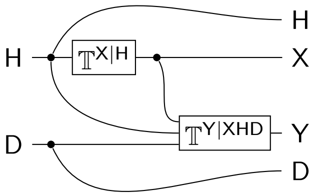

H
H
X
$\mathbb{T}^{\mathsf{X}|\mathsf{HD}}$
$\mathbb{T}^{\mathsf{Y}|\mathsf{XHD}}$
D
Y
D

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

By the assumption $\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$ and version 2 of conditional independence from Theorem 0.1.27,


H
H
X
$\mathbb{T}^{\mathsf{X}|\mathsf{H}}$
*
$\mathbb{T}^{\mathsf{Y}|\mathsf{XHD}}$
D
Y
D

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


H
H
X
$\mathbb{T}^{\mathsf{X}|\mathsf{H}}$
$\mathbb{T}^{\mathsf{Y}|\mathsf{XHD}}$
D
Y
D

$$= \tag{1.4}$$

□

Definition 1.1.3 (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 1.1.4 (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 1.1.5 (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 $\gamma \in \Delta(\mathcal{H})$*

$$\mathbb{U} := (\gamma \otimes \mathrm{Id}^D)\mathbb{T} \tag{1.5}$$

Is a Bayesian see-do model.

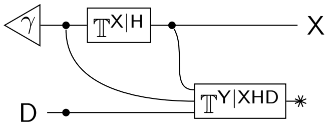
Maybe moves proofs out of main text

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

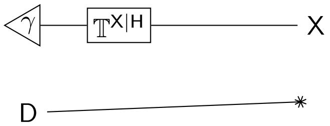
By definition

$$\mathbb{U}^{\mathsf{X}|\mathsf{D}} = \mathbb{U}\mathbb{F}^{\mathsf{X}} \tag{1.6}$$

$$= (\gamma \otimes \mathrm{Id}^D)\mathbb{T}\mathbb{F}^{\mathsf{X}} \tag{1.7}$$

$=$  (1.8)

γ
$\mathbb{T}^{\mathsf{X}|\mathsf{H}}$
X
$\mathbb{T}^{\mathsf{Y}|\mathsf{XHD}}$
D

$=$  (1.9)

Which implies $\mathsf{X} \perp\!\!\!\perp_{\mathbb{U}} \mathsf{D}$ by version (2) of conditional indpendence (Theorem 0.1.27). □

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 hypothesis sufficient see-do model. Define $\mathbb{B} : (0,1) \to \Delta(\{0,1\})$ by $\mathbb{B} : \mathsf{H} \mapsto \mathrm{Bernoulli}(\mathsf{H})$. Then define $\mathbb{T}$ by:

- Choice set: $D = \{0,1\}$
- Observation set: $X = \{0,1\}^N$
- Consequence set: $Y = \{0,1\}$

- Hypothesis set: $H = (0, 1)$
- Observation map: $\mathbb{T}^{\mathsf{X}|\mathsf{H}} : \Upsilon^{N}\mathbb{B}$
- 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

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

$$\begin{aligned}
D_1 := \{ &\text{“walk to the switch and press it with my thumb}'', \\
&\text{“trip over the lego on the floor, hop to the light switch and stab my finger at it}'', \\
&\text{“stay in bed}'' \}
\end{aligned}$$

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:

$$\begin{aligned}D_2 := \{&\text{“walk to the switch and press it with my thumb}'',\\ &\text{“trip over the lego on the floor, hop to the light switch and stab my finger at it}'',\\ &\text{“stay in bed}'',\\ &\text{“toggle the mains power, then flip the light switch}''\}\end{aligned}$$

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 1.1.6 (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 $\mathsf{Y}_2 \perp\!\!\!\perp_{\mathbb{C}} \mathsf{D}|\mathsf{Y}_1\mathsf{H}$.

1.1.2 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 1.1.7 (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 $\mathsf{Y}_1 \not\hookleftarrow_{\mathsf{Y}_2} \mathsf{D}$

Lemma 1.1.8 (Limited unresponsiveness implies D-causation)**.** *For deterministic consequence models,* $\mathsf{Y}_1 \not\hookleftarrow_{\mathsf{Y}_2} \mathsf{D}$ *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 $A \in \mathcal{Y}_2$, $B \in \mathcal{Y}_1$, $s \in S$, $d \in D$:

$$\begin{aligned}
\mathbb{C}^{\mathsf{Y}_2|\mathsf{SD}}_{(s,d)}\curlyvee(\mathbb{M}\otimes\mathrm{Id}) &= \int_A \mathbb{M}(y_2', d, s; B) d\delta_{y_2(s,d)}(y_2') && (1.10)\\
&= \int_A \delta_{f(y_2',s)}(B) d\delta_{y_2(s,d)}(y_2') && (1.11)\\
&= \delta_{f(y_2(s,d),s)}(B)\delta_{y_2(s,d)}(A) && (1.12)\\
&= \delta_{y_1(s,d)}(B)\delta_{y_2(s,d)}(A) && (1.13)\\
&= \delta_{y_2(s,d)}\otimes\delta_{y_1(s,d)}(A\times B) && (1.14)
\end{aligned}$$

$\mathbb{M}$ 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

$$\mathbb{C}^{\mathsf{LF}|\mathsf{DS}}_{d,[i,m]} = \delta_{(d \text{ XOR } i) \text{ AND } m}\otimes\delta_{(d \text{ XOR } i) \text{ AND } m} \qquad (1.15)$$

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

$$\mathbb{C}_\gamma = \frac{1}{4}\mathbb{C}_{\cdot,[0,0]} + \frac{1}{4}\mathbb{C}_{\cdot,[0,1]} + \frac{1}{2}\mathbb{C}_{\cdot,[1,1]} \qquad (1.16)$$

$$\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma = \frac{1}{4}\begin{bmatrix}1&0\\0&1\end{bmatrix}\otimes\begin{bmatrix}1&0\\1&0\end{bmatrix} + \frac{1}{4}\begin{bmatrix}1&0\\0&1\end{bmatrix}\otimes\begin{bmatrix}1&0\\0&1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}0&1\\1&0\end{bmatrix}\otimes\begin{bmatrix}0&1\\1&0\end{bmatrix} \qquad (1.17)$$

Then

$$\begin{aligned}
[1,0]\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma &= \frac{1}{4}[0,1]\otimes[1,0] + \frac{1}{4}[0,1]\otimes[0,1] + \frac{1}{2}[1,0]\otimes[1,0] && (1.18)\\
[1,0]\curlyvee(\mathbb{C}^{\mathsf{F}|\mathsf{D}}_\gamma\otimes\mathbb{C}^{\mathsf{L}|\mathsf{D}}_\gamma) &= (\frac{1}{2}[0,1] + \frac{1}{2}[1,0])\otimes(\frac{1}{4}[0,1] + \frac{3}{4}[1,0]) && (1.19)\\
\implies [1,0]\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma &\neq [1,0]\curlyvee(\mathbb{C}^{\mathsf{F}|\mathsf{D}}_\gamma\otimes\mathbb{C}^{\mathsf{L}|\mathsf{D}}_\gamma) && (1.20)
\end{aligned}$$

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

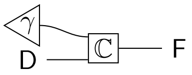
define this

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

Theorem 1.1.9 shows that under some conditions, D-causation can hold for arbitrary mixtures over subsets of the hypothesis class H.

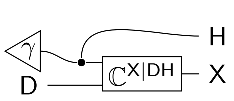
Theorem 1.1.9 (Universal D-causation)**.** *If* $\mathsf{X} \perp\!\!\!\perp \mathsf{H}|\mathsf{D}$ *for all* $\mathsf{H}, \mathsf{H}' \in S \subset \mathsf{H}$ *and* X *D-causes* Y *in all* $\mathsf{H} \in S$*, then* X *D-causes* Y *with respect to all mixed consequence models* $\mathbb{C}_\gamma$ *for all* $\gamma \in \Delta(\mathsf{H})$ *with* $\gamma(S) = 1$.

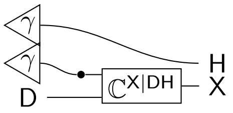
Proof. For $\gamma \in \Delta(\mathsf{H})$, define the mixture

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


(1.21)

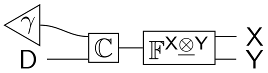
Because $\mathbb{C}_{\mathsf{H}}^{\mathsf{X}|\mathsf{D}} = \mathbb{C}_{\mathsf{H}'}^{\mathsf{X}|\mathsf{D}}$ for all $\mathsf{H}, \mathsf{H}' \in \mathsf{H}$, we have



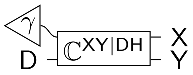
$=$


(1.22)

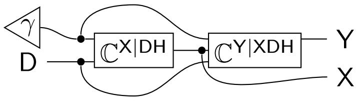
Also

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


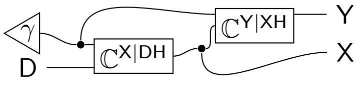
(1.23)

$=$


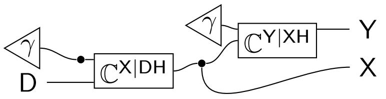
(1.24)

$=$


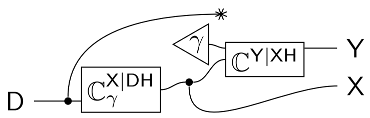
(1.25)

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


(1.26)

$\overset{1.22}{=}$


(1.27)

$= \overset{1.22}{=}$


(1.28)

Equation 1.28 establishes that $(\gamma \otimes \mathbf{Id}_X \otimes \overset{*}{\mid}_D)\mathbb{C}^{\mathsf{Y|XH}}$ is a version of $\mathbb{C}_\gamma^{\mathsf{Y|XD}}$, and thus $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{C}_\gamma} \mathsf{D|X}$.

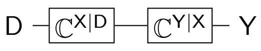
This can also be derived from the semi-graphoid rules:

$$\begin{aligned}
\mathsf{H} \perp\!\!\!\perp \mathsf{D} \wedge \mathsf{H} \perp\!\!\!\perp \mathsf{X|D} &\implies \mathsf{H} \perp\!\!\!\perp \mathsf{XD} && (1.29)\\
&\implies \mathsf{H} \perp\!\!\!\perp \mathsf{D|X} && (1.30)\\
\mathsf{D} \perp\!\!\!\perp \mathsf{H|X} \wedge \mathsf{D} \perp\!\!\!\perp \mathsf{Y|XH} &\implies \mathsf{D} \perp\!\!\!\perp \mathsf{Y|X} && (1.31)\\
&\implies \mathsf{Y} \perp\!\!\!\perp \mathsf{D|X} && (1.32)
\end{aligned}$$

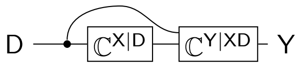
□

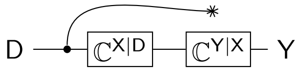
1.1.3 Properties of D-causation

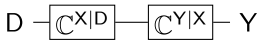
If X D-causes Y relative to $\mathbb{C}_\mathsf{H}$, then the following holds:

$\mathbb{C}_\mathsf{H}^{\mathsf{X|D}} =$  (1.33)

This follows from version (2) of Definition 0.1.29:

$\mathbb{C}_\mathsf{H}^{\mathsf{X|D}} =$  (1.34)

$=$  (1.35)

$=$  (1.36)

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?

1.1.4 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 not requirement that D_1 comes "before" D_2 or even that they have any "before" and "after" relations at all.

Define parallel decisions, show that they induce potential outcomes

1.1.5 Residual dependence on observations

Definition 1.1.10 (Hypothesis sufficiency)**.** The hypothesis H is *sufficient* for a see-do model if the consequence model has no dependence on observations X conditional on H. That is, $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{X}|\mathsf{D}\mathsf{H}$.

A hypothesis sufficient see-do model can be specified with:

- Hypothesis space H, choices D, observations X and consequences Y
- Observation map $\mathbb{T}^{\mathsf{X}|\mathsf{H}}$
- Reduced consequence model $\mathbb{T}^{\mathsf{Y}|\mathsf{HD}}$

Given observations X, assumed to be an IID sequence $\mathsf{X}_1, \mathsf{X}_2, ...$ conditional on H, a common "causal inference problem" is to estimate the "true" distribution of observations $\mathbb{T}^{\mathsf{X}_i|\mathsf{H}}_{h^*}$ and from this to estimate the consequence model $\mathbb{T}^{\mathsf{Y}|\mathsf{HD}}_{h^*\cdot}$, if this is possible. This problem only makes sense if hypothesis sufficiency is assumed – once h^* is given, the consequence model of interest has no further dependence on X. We show that all decision problems can be modeled by a hypothesis sufficient see-do model.

Examples of hypothesis sufficient and insufficient 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 hypothesis sufficient 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:

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

In this model, the chance H of the coin landing on heads is as much as we can hope to know about 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 hypothesis insufficient 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 appears that it might be possible to model the second situation with a hypothesis sufficient model by including the result of the $N+1$th flip in the hypothesis. Define the new hypothesis space $H' = (0, 1) \times \{0, 1\}$ and define ${}^{3}\mathbb{T}$ by:

- $D = \{0, 1\}$
- $X = \{0, 1\}^{N+1}$
- $Y = \{0, 1\}$
- $H' = (0, 1) \times \{0, 1\}$
- ${}^{3}\mathbb{T}^{\mathsf{X}|\mathsf{H}'} : (\curlyvee^{N}\mathbb{B} \otimes \delta_{x_{N+1}}$
- ${}^{3}\mathbb{T}^{\mathsf{Y}|\mathsf{H}'\mathsf{D}} : (h, x_{N+1}, d) \mapsto \delta_{1-|d-x_{N+1}|}$

However, X_{N+1} is related to the previous flips $\boldsymbol{\mathsf{X}}_{<N}$ and ${}^{3}\mathbb{T}$ ignores this fact. In particular, given any $\mathsf{H}' = (h, \_)$, X_{N+1} as well as X_i, $i \leq N$ should all distributed according to Bernoulli(h). Thus ${}^{2}\mathbb{T}$ is preferable to ${}^{3}\mathbb{T}$ because it represents more of the knowledge we have about the problem.

If a see-do model is employed in a *decision problem* – defined in the next section – there is an alternative way to avoid hypothesis insufficiency that does not require throwing out some of the model structure.

The importance of this is that counterfactual questions are usually *not* decision problems and so they do not have the possibility of avoiding insufficiency available; also *in practice* counterfactual problems are usually hypothesis insufficient while decision problems are usually not.

1.1.6 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* $\mathbb{D} : X \to \Delta(\mathcal{D})$.

Avoiding insufficiency with decision functions

Show that a decision problem with a hypothesis insufficient model induces an equivalent decision problem with a hypothesis sufficient 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 1.1.11 (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 1.1.12 (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 1.1.13 (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 1.1.14 (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

$$R(\mathsf{H}, \mathbb{U}) := -\int_X \mathbb{U}_x \mathbb{T}^{\mathsf{Y|DXH}}_{\cdot,x,\mathsf{H}} u d\mathbb{T}^{\mathsf{X|H}}_{\mathsf{H}}(x) \tag{1.37}$$

for $\mathsf{H} \in \mathsf{H}$, $\mathbb{U} \in \mathcal{U}$. Here $\mathbb{U}_x \mathbb{T}^{\mathsf{Y|DXH}}_{\cdot,x,\mathsf{H}} u$ is the product of the measure $\mathbb{U}_x$, the kernel $\mathbb{T}^{\mathsf{Y|DXH}}_{\cdot,x,\mathsf{H}} : D \to \Delta(\mathcal{Y})$ 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 1.1.15 (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 hypothesis sufficient 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 $Y_{\cdot}(\cdot) : D \times U \to Y$ (distinct from Y, which appears to be a function $U \to$ something) and calls $Y_{\cdot}(\cdot)$ the "potential response". We could always consider the variable $\mathsf{V} := \underline{\otimes}_{i \in [n]} \mathsf{V}_i$ and define the "total potential response" $\mathbf{g} := \mathsf{V}_{\cdot}(\cdot)$, which captures the potential responses of any subset of variables in V.

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

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

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

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

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

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

One obvious question is: does $\mathbf{g}$ have to be deterministic? While SCMs are defined in terms of deterministic functions with noise arguments, it's not clear that this is a necessary feature of counterfactual models. If $\mathbf{g}$ were properly stochastic, what is the problem with considering $\mathbf{g}(x, u)$ and $\mathbf{g}(x', u)$ to represents what would happen in a fixed situation u if I did x and if I did x' respectively? In fact, a nondeterministic $\mathbf{g}$ arguably fails to capture a key intuition of taking actions "in exactly the same situation". If I want to know the result of doing action x and, in exactly the same situation, the result of doing action x, then one might intuitively think that the result should always be *deterministically the same.* This property, which we call *deterministic reproducibility*, does not hold if we consider a nondeterministic potential response map $\mathbf{g}$.

This idea of doing x and, in the same situation, doing x doesn't render very well in English. Furthermore, even though deterministic reproducibility seems to be an important property of counterfactual SCMs, they don't help very much to elucidate the idea. "If I take action x in situation U I get $V_x(u)$ and if I take action x in situation U I get $V_x(u)$" is just a redundant repetition. It seems that we want some way to express the idea of having two copies of $V_x(u)$ or, more generally, having multiple copies of a potential response function in such a way that we can make comparisons between their results.

The idea that we need *can* be clearly expressed with a see-do model.

1.3 Functional Exchangeability, Imitability, Potential Outcomes

A great deal of "non-causal statistics" deals with models of *exchangeable random variables.* Random variables X_1 and X_2 both taking values in X are exchangeable if $\mathbb{P}^{\mathsf{X}_1\mathsf{X}_2} = \mathbb{P}^{\mathsf{X}_2\mathsf{X}_1}$; more generally, a set of n random variables is exchangeable if the joint distribution is unchanged by applying any permutation to the order of the variables. Exchangeability implies indifference to the ordering of *observations.* For example:

- Suppose the random variables $\mathsf{X}_1, \mathsf{X}_2, \mathsf{X}_3$ represent beliefs about the outcomes of flipping a coin twice. If we are not completely sure the coin is fair, then we might suppose that seeing heads for the first two flips will leave us slightly more confident in a heads than a tails outcome on the next flip, so $\mathbb{P}^{\mathsf{X}_3|\mathsf{X}_2\mathsf{X}_1}_{\mathrm{HH}}$ is not quite the same as $\mathbb{P}^{\mathsf{X}_3}$. However, if $\mathsf{X}_1, \mathsf{X}_2$ and X_3 are exchangeable then it must be the case that $\mathbb{P}^{\mathsf{X}_3|\mathsf{X}_2\mathsf{X}_1}_{HT} = \mathbb{P}^{\mathsf{X}_3|\mathsf{X}_1\mathsf{X}_2}_{HT}$; our views about the likelihood of a particular X_3 after seeing a particular number of heads and a particular number of tails do not depend on the order of the heads and tails.

- Suppose W_1 and W_2 represent consecutive draws without replacement from an urn containing a finite number of red and white balls in the frequentist sense that given a large ensemble of draws from the urn, the sample fractions will converge to the respective probabilities of W_1 and W_2. While the changing number of balls in the urn makes W_2 depend on W_1, it can be verified that supposing an equal chance of drawing any particular ball, W_1 and W_2 are exchangeable, i.e. $\mathbb{P}^{\mathsf{W}_1\mathsf{W}_2} = \mathbb{P}^{\mathsf{W}_2\mathsf{W}_1}$. This means that the probability of drawing a certain number of reds and whites does not depend on the order in which the balls are drawn.

Functional exchangeability is a causal version of exchangeability that applies to consequence maps. Where exchangeability implies indifference to the ordering of observations, functional exchangeability implies indifference to the ordering of *decision-consequence pairs*. For example:

- Suppose you are planning to conduct a sequence of experiments in which you will set a piece of equipment to a particular setting, represented by $\mathsf{D}_1, \mathsf{D}_2, \mathsf{D}_3$, and record the results. The random variables $\mathsf{Y}_1, \mathsf{Y}_2, \mathsf{Y}_3$ represent our beliefs about the results of the first, second and third repeats of the experiment prior to undertaking it. This experiment is functionally exchangeable if it is described by a consequence map $\mathbb{C} : D_0^3 \to \Delta(\mathcal{Y}^3)$ such that $\mathbb{C}^{\mathsf{Y}_1\mathsf{Y}_2\mathsf{Y}_3|\mathsf{D}_1\mathsf{D}_2\mathsf{D}_3} = \mathbb{C}^{\mathsf{Y}_2\mathsf{Y}_3\mathsf{Y}_1|\mathsf{D}_2\mathsf{D}_3\mathsf{D}_1}$ and likewise for other joint permutations of the decisions and consequences. As in the coin flipping example, this means that we "learn the same thing" from choosing settings a and observing p followed by choosing b and observing q as we would from choosing b and observing q followed by choosing a and observing p

- Suppose W_1 and W_2 represent (in the frequentist sense) an ensemble of draws from one of two finite urns of red and white balls, with the urn choices given by D_1, D_2. Then $\mathsf{W}_1, \mathsf{W}_2$ and $\mathsf{D}_1, \mathsf{D}_2$ are conditionally exchangeable in the sense that $\mathbb{P}^{\mathsf{W}_1\mathsf{W}_2|\mathsf{D}_1\mathsf{D}_2} = \mathbb{P}^{\mathsf{W}_2\mathsf{W}_1|\mathsf{D}_2\mathsf{D}_1}$

Define permutation, set permutation

Definition 1.3.1 (Exchangeability)**.** A (possibly infinite) sequence of random variables $\mathsf{X}_A := \underline{\bigotimes}_{i\in A}\mathsf{X}_i$ taking values in X on a probability space $(\mathbb{P}, (\Omega, \mathcal{F}))$ is *exchangeable* if for all $n \in A$, all permutations of indices $\sigma : [n] \to [n]$ and all $B \in X^n$ such that $B\times \in \mathcal{F}$, $\mathbb{P}^{\mathsf{X}_{[n]}\mathsf{X}_{A\setminus[n]}}(B \times X^{|A|-n}) = \mathbb{P}^{\mathsf{X}_{\sigma([n])}\mathsf{X}_{A\setminus[n]}}(B \times X^{|A|-n})$.

Define infinite coupled copymaps WRT kolmogorov extension theorem

De Finetti's representation theorem shows that any infinite exchangeable sequence of random variables can be represented as a mixture of independent and identically distributed sequences. This theorem connects exchangeable *subjective Bayesian random variables* $\mathsf{X}_n^{(B)}$, which represent expectations of future X^n-valued observations, and independent and identically distributed *frequentist random variables* $\mathsf{X}^{(F)}$, which represent almost sure $n \to \infty$ limits of proportions in X^n-valued observations. Conseuqences of the theorem are:

- Any exchangeable sequence of random variables X_A that can be extended to an infinite sequence can be represented as a mixture $\mu \in \Delta(\theta)$ of coupled statistical models $\underline{\bigotimes}_{i\in A}\mathbb{O}$ where $\mathbb{O} : \Theta \to \Delta(\mathcal{X})$

- Any statistical model $\mathbb{O} : \Theta \to \Delta(\mathcal{X})$ along with a prior μ induces a family of exchangeable sequences of random variables X_A, $A \subset \mathbb{N}$

If our task is to predict future X-valued observations and our predictions are the same regardless of the order in which the observations arrive, then the probability distribution representing predictions can be represented by a statistical model and a prior. A similar theorem exists connects "functionally exchangeable" consequence models and coupled 2-player statistical models. Coupled 2-player statistical models themselves are closely related to models typically used in causal inference.

Theorem 1.3.2 (Representation of infinite exchangeable sequences(Hewitt and Savage, 1955))**.** *Let $(X, \mathcal{X})$ be a compact Hausdorff space with the Baire σ-algebra and $\mathbb{P}$ a measure on $X^{\mathbb{N}}$ with the product sigma algebra such that $\mathsf{X}_{\mathbb{N}}$ with each X_i given by the projection map π_i. Define $(\Delta(\mathcal{X}), \mathcal{E})$ to be the set of all probability measures on X with the σ-algebra $\mathcal{E}$ being the coarsest algebra for which the maps $\mathrm{ev}_A : \Delta(\mathcal{X}) \to \mathbb{R}$ given by $\mathrm{ev}_A : \rho \mapsto \rho(A)$ are measurable for all $A \in \mathcal{X}$. Then there exists a unique probability measure μ on $\Delta(\mathcal{X})$ with the given σ-algebra such that for all $n \in \mathbb{N}$, $C \in \mathcal{X}^{\setminus}$:*

$$\mathbb{P}(C \times X^{\mathbb{N}\setminus[n]}) = \int_{(\Delta(\mathcal{X}))} \prod_{i \in [n]} \rho(\mathsf{X}_i(C)) d\mu(\rho) \tag{1.38}$$

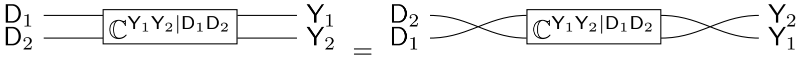
Where $\mathsf{X}_i(A)$ is the image of A under X_i.

Examples of such spaces include bounded Borel subsets of $\mathbb{R}$ and bounded subsets of $\mathbb{N}$.

Definition 1.3.3 (Functional Exchangeability)**.** A finite sequence of random variables $\mathsf{Y}_1, ..., \mathsf{Y}_n : \Omega \to Y$ 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

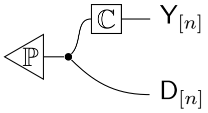


(1.39)

Include a lemma about swap maps and variable permutations

Lemma 1.3.4 (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:*

$\mathbb{P}' =$



(1.40)

Then sequence $\bigotimes_{i \in [n]}(\mathsf{Y}_i \otimes \mathsf{D}_i)$ (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

$$\mathbb{P}'^{\mathsf{X}_1\mathsf{D}_1,\ldots,\mathsf{X}_n\mathsf{D}_n}(\prod_{i\in[n]} A_i \times B_i) = \int_{\prod_{i\in[n]} B_i} \mathbb{P}'^{\mathsf{X}_{[n]}|\mathsf{D}_{[n]}}_{d_{[n]}}(\prod_{i\in[n]} A_i) d\mathbb{P}'^{\mathsf{D}_{[n]}}(d_{[n]}) \tag{1.41}$$

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

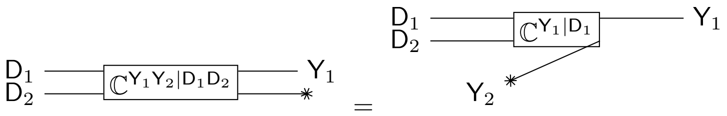
$$= \int_{\prod_{i\in[n]} B_i} \mathbb{C}^{\mathsf{X}_{\sigma([n])}|\mathsf{D}_{\sigma([n])}}_{d_{[n]}}(\prod_{i\in[n]} A_i) d\mathbb{P}^{\mathsf{D}_{\sigma([n])}}(d_{[n]}) \tag{1.43}$$

$$= \int_{\prod_{i\in[n]} B_i} \mathbb{P}'^{\mathsf{X}_{\sigma([n])}|\mathsf{D}_{\sigma([n])}}_{d_{[n]}}(\prod_{i\in[n]} A_i) d\mathbb{P}'^{\mathsf{D}_{\sigma([n])}} \tag{1.44}$$

$$= \mathbb{P}'^{\mathsf{X}_{\sigma(1)}\mathsf{D}_{\sigma(1)},\ldots,\mathsf{X}_{\sigma(n)}\mathsf{D}_{\sigma(n)}}(\prod_{i\in[n]} A_i \times B_i) \tag{1.45}$$

Where line 1.43 follows from exchangeability of $\mathbb{P}$ and functional exchangeability of $\mathbb{C}$ and lines 1.42 and 1.44 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 1.3.5 (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 $\mathbb{C}^{\mathsf{Y}_U|\mathsf{D}_{[n]}} = \mathbb{C}^{\mathsf{Y}_U|\mathsf{D}_U}$. Noninterference implies that discard maps can "fall through" kernels:


D₁
D₂
$\mathbb{C}^{\mathsf{Y}_1\mathsf{Y}_2|\mathsf{D}_1\mathsf{D}_2}$
Y₁
=
D₁
D₂
$\mathbb{C}^{\mathsf{Y}_1|\mathsf{D}_1}$
Y₁
Y₂

(1.46)

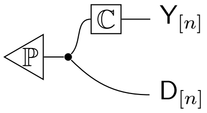
I think it is the case that functionally exchangeable + non-interfering implies infinitely functionally exchangeably extendable, but not proved yet

Theorem 1.3.6 (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 $n \in \mathbb{N}$, $d \in D^n$, $C \in \mathcal{Y}^n$:*

$$\mathbb{C}_d(C) = \int_F \prod_{i\in[n]} \mathbb{H}_{\mathsf{D}_i(d)}(\mathsf{Y}_i(C)) d\nu(\mathbb{H}) \tag{1.47}$$

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

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



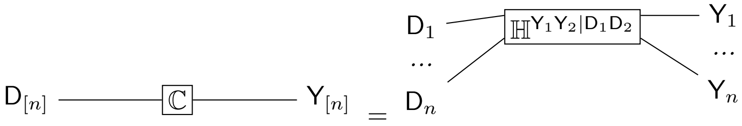
(1.48)

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

. By Lemma 1.3.4, there is an exchangeable sequence $\underline{\bigotimes}_{i\in[n]}\mathsf{Y}'_i\underline{\otimes}\mathsf{D}'_i$ 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 1.3.7. *Equivalently, for all $n \in \mathbb{N}$, let $\mathbb{G} : F \to$*



(1.49)

Chapter 2

Chapter 4: See-do models compared to causal graphical models and potential outcomes

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.

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.

Edwin Hewitt and Leonard J. Savage. Symmetric Measures on Cartesian Products. *Transactions of the American Mathematical Society*, 80(2):470–501, 1955. ISSN 0002-9947. doi: 10.2307/1992999. URL https://www.jstor.org/stable/1992999. Publisher: American Mathematical Society.

Alan Hájek. What Conditional Probability Could Not Be. *Synthese*, 137(3): 273–323, December 2003. ISSN 1573-0964. doi: 10.1023/B:SYNT.0000004904.91112.16. URL https://doi.org/10.1023/B:SYNT.0000004904.91112.16.

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.

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`.

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: